



<http://researchspace.auckland.ac.nz>

ResearchSpace@Auckland

Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage.

<http://researchspace.auckland.ac.nz/feedback>

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form.

HOW DO PEOPLE MANAGE THEIR DOCUMENTS?

**AN EMPIRICAL INVESTIGATION INTO
PERSONAL DOCUMENT MANAGEMENT PRACTICES
AMONG KNOWLEDGE WORKERS**

SARAH HENDERSON

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy,
The University of Auckland, 2009

Personal document management is the activity of managing a collection of digital documents performed by the owner of the documents, and consists of creation/acquisition, organisation, finding and maintenance. Document management is a pervasive aspect of digital work, but has received relatively little attention from researchers. The hierarchical file system used by most people to manage their documents has not conceptually changed in decades. Although revolutionary prototypes have been developed, these have not been grounded in a thorough understanding of document management behaviour and therefore have not resulted in significant changes to document management interfaces.

Improvements in understanding document management can result in productivity gains for knowledge workers, and since document management is such a common activity, small improvements can deliver large gains. The aim of this research was to understand how people manage their personal document collections and to develop guidelines for the development of tools to support personal document management.

A field study was conducted that included interviews, a survey and file system snapshot. The interviews were conducted with ten participants to investigate their document management strategies, structures and struggles. In addition to qualitative analysis of semi-structured interviews, a novel investigation technique was developed in the form of a file system snapshot which collects information about document structures and derives a number of metrics which describe the document structure. A survey was also conducted, consisting of a questionnaire and a file system snapshot, which enabled the findings of the field study to be validated, and to collect information from a greater number of participants.

The results of this research culminated in (1) development of a conceptual framework highlighting the key personal document management attitudes, behaviours and concerns; (2) model of basic operations that any document management system needs to provide; (3) identification of piling, filing and structuring as three key document management strategies; (4) guidelines for the development of user interfaces to support document management, including specific guidelines for each document management strategy. These contributions both improve knowledge of personal document management on which future research can build, and provide practical advice to document management system designers which should result in the development of more usable system.

ACKNOWLEDGEMENTS

This thesis is dedicated to the memory of Moeroa Butland. Moeroa was a wonderful and patient friend who was always there for me when I needed someone.

I would like to thank my supervisor, Professor Ananth Srinivasan, and my advisors, Professor Michael Myers and Associate Professor David Sundaram for being so patient and supportive. I would also like to thank everyone else in the Department of Information Systems and Operations Management who encouraged me during this process, and particularly Associate Professor Don Sheridan, for all his help and support.

After thinking I was the only person interested in personal information management, I was delighted to meet some fellow researchers in this area: Rick Boardman, Ofer Bergman, Daniel Gonçalves and William Jones.

Thanks are due to all the people who nagged me about my thesis: Mark, Angela and Hemant. Thanks are also due to those who didn't: Andrew, Jo, Doug, Raina, Amal, Grant and Koro, as well as my family: Johan and Paul, Nardia and Mark, and Kyle.

This research was financially supported by a Top Achiever Doctoral Scholarship from FRST, and a Doctoral Scholarship from the University of Auckland, for which I am extremely grateful.

And finally, Stephen Witherden deserves endless praise for being eternally patient and encouraging, listening to me talk about this topic for years on end, giving lots of constructive advice and commenting on many versions of this thesis. I absolutely could not have survived this process without him.

TABLE OF CONTENTS

INTRODUCTION	1
1.1 Research Problem	1
1.2 Aim	3
1.3 Contributions	3
1.4 Research Approach	4
LITERATURE REVIEW	5
2.1 Terms and Definitions	5
2.2 Empirical studies of Personal Information Management	11
2.2.1 Paper documents	12
2.2.2 Email	14
2.2.3 Web bookmarks	15
2.2.4 Digital Documents	15
2.3 Prototypes and Systems	18
2.3.1 Personal Document Management System History	19
2.3.2 Windows Explorer (in Windows XP)	20
2.3.3 'Pile' User Interface	23
2.3.4 Lifestreams	25
2.3.5 TimeScape	26
2.3.6 Presto	28
2.3.7 Haystack	29
2.3.8 Stuff I've Seen	30
2.3.9 Google Desktop	31
2.3.10 Copernic Desktop Search	31
2.4 Theory Related To Personal Document Management	33
2.4.1 Models of Personal Information Management	33
2.4.2 Theories of Workspaces and Distributed Cognition	34
2.4.3 Theories of Classification	35
2.4.4 Theories of Information Retrieval	39
2.5 Conclusion	41
RESEARCH DESIGN	43
3.1 Research Questions	43
3.2 Type of Research	44
3.3 Research Strategy	44
3.4 Data Collection Methods	47
3.4.1 Interviews and File System Snapshot	48
3.4.2 Survey	49
3.5 Participant Selection	49
3.5.1 Generalisability	50
3.6 Ethical Considerations	51
3.7 Conclusion	51
INTERVIEWS AND FILE SYSTEM SNAPSHOT	53
4.1 Study Design	53
4.1.1 Choice of Method	54
4.1.2 Participants	54
4.1.3 Interview Process	55
4.2 Interview Results	57
4.2.1 Interview Summaries	58
4.2.2 Interview Analysis	67
4.2.3 Summary of Interview Analysis	104

4.3	File System Snapshot	108
4.3.1	Metrics	110
4.3.2	Data Cleansing	111
4.3.3	Analysis Software	111
4.3.4	Snapshot Results.....	113
4.3.5	Folder Name Analysis	122
4.4	Document Management Strategies.....	133
4.4.1	Piling	133
4.4.2	Filing	133
4.4.3	Structuring	134
4.5	Conclusion.....	134
SURVEY	137
5.1	Survey Design.....	137
5.1.1	Questionnaire Design	142
5.1.2	Testing	154
5.2	Survey Results	157
5.2.1	Questionnaire Section 1: Attitudes	158
5.2.2	Survey Section 2: New System Features.....	160
5.2.3	Survey Section 3: Desktop	165
5.2.4	Survey Section 4: My Documents	169
5.2.5	Survey Section 5: Creating and Naming Files	170
5.2.6	Survey Section 6: Creating and Naming Folders.....	173
5.2.7	Survey Section 7: Locating Documents.....	174
5.2.8	Survey Section 8: Searching.....	178
5.2.9	Survey Section 9: Viewing Documents	181
5.2.10	Survey Section 10: Versions.....	185
5.2.11	Survey Section 11: Copies.....	187
5.2.12	Survey Section 12: Deleting and Backup	188
5.2.13	Survey Section 13: Demographics	191
5.2.14	Survey Section 14: Comments	195
5.3	File System Snapshot Results.....	196
5.3.1	Overall Size	197
5.3.2	Locations.....	198
5.3.3	Depth	200
5.3.4	Bushiness	201
5.3.5	Leafiness	206
5.3.6	Balance	207
5.3.7	Emptiness	209
5.3.8	Duplication	210
5.3.9	Shortcuts.....	213
5.3.10	File Types	214
5.3.11	File Names	217
5.3.12	Folder Names.....	218
5.3.13	Versions	218
5.3.14	Satisfaction vs. File System Metrics.....	219
5.3.15	Academic and General staff.....	219
5.4	Document Management Strategies.....	222
5.5	Conclusion.....	223
DISCUSSION	225
6.1	Conceptual Model validation and refinement.....	225
6.1.2	Final Conceptual Model.....	234
6.2	Document Management System Capabilities.....	234
6.3	Document Management Strategies.....	236
6.3.1	Piling	238

6.3.2 Filing.....	239
6.3.3 Structuring	239
6.3.4 Relationship between structure and strategy	239
6.4 User Interface Guidelines	240
6.4.1 Personas.....	241
6.4.2 General User Interface Guidelines	250
6.5 Summary.....	254
CONCLUSION	257
7.1 Summary of Research Problem and Approach.....	257
7.2 Contributions	259
7.2.1 Contributions to understanding personal document management	259
7.2.2 Contributions to user interface guidelines	260
7.3 Limitations And Future Work.....	261
REFERENCES	263
APPENDICES	270

TABLE OF FIGURES

Figure 1: Conceptualisation of Personal Information Management as consisting of acquisition, organisation, retrieval and maintenance (Boardman, 2004, p. 17).....	6
Figure 2: Relationship of Information Retrieval to Personal Information Management (Boardman, 2004) ...	11
Figure 3: Task-Artifact Cycle (from Carroll et al., 1991).....	12
Figure 4: Personal computer operating system market share (NetApplications Ltd, 2008).....	21
Figure 5: Microsoft Windows Explorer interface (Windows XP version).....	22
Figure 6: Files shown using Windows XP Details view.....	22
Figure 7: Files shown using Windows XP List view	23
Figure 8: Files shown using Windows XP Icons view	23
Figure 9: Pile interface showing (a) user-created pile and (b) system-created pile (Mander et al., 1992)	24
Figure 10: Lifestreams interface (Fertig et al., 1996b).....	26
Figure 11: TimeScape interface showing static desktop (Rekimoto, 1999a).....	27
Figure 12: TimeScape interface showing timelines (Rekimoto, 1999a).....	27
Figure 13: The Vista Interface for the Presto system (Dourish et al., 1999a)	28
Figure 14: Haystack user interface (Karger et al., 2003).....	29
Figure 15: Stuff I've Seen 'Top View' user interface (Dumais et al., 2003)	30
Figure 16: Google Desktop search results	31
Figure 17: Copernic Desktop Search interface.....	32
Figure 18: Example Folder Structure	37
Figure 19: Trade-off between filing and finding	40
Figure 20: Trade-off between filing and finding shifting with improved search	41
Figure 21: Overall field study research strategy combining interviews and survey	47
Figure 22: Conceptual model created from interview analysis	68
Figure 23: Desktop screenshots of seven participants	91
Figure 24: Interface of File System Snapshot Software	109
Figure 25: Information Structure Analyser software interface.....	112
Figure 26: Bar graph showing number of files for each participant	113
Figure 27: Bar graph showing number of folders for each participant.....	113
Figure 28: Bar graph showing unique files and total files for each participant	114
Figure 29: Bar graph showing empty folders and total folders for each participant.....	114
Figure 30: Bar graph showing average and maximum depth for each participant	115
Figure 31: Bar graph showing bushiness - average number of subfolders per folder	115
Figure 32: Bar graph showing standard deviation of subfolders per folder	116
Figure 33: Bar graph showing maximum number of subfolders per folder.....	116
Figure 34: Bar graph showing branching factor for each participant	117
Figure 35: Bar graph showing leafiness - average number of files per folder	117
Figure 36: Bar graph showing standard deviation of files per folder	118
Figure 37: Bar graph showing maximum number of files per folder.....	118
Figure 38: Bar graph showing number of shortcuts	119
Figure 39: Bar graph showing the number of shortcuts in each location.....	119
Figure 40: Bar graph showing the number of folders in each location.....	120
Figure 41: Bar graph showing the number of files in each location	120
Figure 42: Bar graph showing the percentage of file and folder duplication	121
Figure 43: Bar graph showing percentage of folders in top level	121
Figure 44: Bar graph showing percentage of files in top level.....	122
Figure 45: Average Proportion of Folders with each Code	124
Figure 46: Radial Graphs showing profile of the top four codes	128
Figure 47: Three equivalent folder structures	132
Figure 48: First Contact - Invitation email sent to survey participants.....	145
Figure 49: Second Contact - Introduction page on survey website	146
Figure 50: Third Contact - Combination of thank you and reminder email sent to participants one week after the initial contact.	148
Figure 51: Survey page showing first two survey questions.....	149
Figure 52: Survey question showing skip link and disabled question depending on user response.	150
Figure 53: Survey progress indicators at three different times during the survey, (a) at the start, (b) partway through and (c) at the end.....	151

Figure 54: File System Snapshot page of the questionnaire.....	152
Figure 55: File System Snapshot software waiting to run.....	153
Figure 56: File System Snapshot software after running.....	153
Figure 57: Questionnaire Thank You page.....	154
Figure 58: Question 1 - Bar graph showing how organised participants describe themselves as.....	158
Figure 59: Normal Q-Q Plot of Satisfaction Factor Analysis Score.....	160
Figure 60: Question 6 - Bar chart showing what participants use as a Desktop background.....	165
Figure 61: Question 7 - Bar chart showing whether participants keep documents on their Desktop.....	165
Figure 62: Box plot showing satisfaction is lower if documents are kept on the Desktop.....	165
Figure 63: Question 8 - Bar chart showing whether Desktop documents are organised into folders.....	166
Figure 64: Question 9 - Bar chart showing who created the documents on the Desktop.....	166
Figure 65: Question 10 - Bar chart showing whether participants use the spatial abilities of the Desktop..	166
Figure 66: Question 11 - Bar chart showing how long documents tend to stay on the Desktop.....	167
Figure 67: Question 12 - Bar chart showing the main purpose of documents on the Desktop.....	167
Figure 68: Question 13 - Bar chart showing where a document goes after it leaves the Desktop.....	168
Figure 69: Question 14 - Bar chart showing the main reason for putting documents on the Desktop.....	168
Figure 70: Question 15 - Bar chart showing whether the participants use the My Documents folder to store documents.....	169
Figure 71: Question 16 - Bar chart showing why participants don't use the My Documents folder.....	169
Figure 72: Question 17 - Bar chart showing how new documents are created.....	170
Figure 73: Question 18 - Bar chart showing when documents are usually named.....	170
Figure 74: Question 19 - Bar chart showing whether other people's files get renamed.....	171
Figure 75: Question 20 - Bar chart showing how often document naming schemes are used.....	171
Figure 76: Question 21 - Bar chart showing whether the folder name is included in the file name.....	171
Figure 77: Question 23 - Bar chart showing minimum information needed to identify a document's contents.....	172
Figure 78: Question 24 - Bar graph showing when participants create folders.....	173
Figure 79: Box plot showing satisfaction is lower if folders are created after files.....	173
Figure 80: Question 25 - Bar graph showing the number of people who rated each folder naming item as either 'important or 'very important'.....	174
Figure 81: Question 26 - Bar graph showing find method for a recently-used document.....	175
Figure 82: Box plot showing how satisfaction differs by find method for a recently-used document.....	175
Figure 83: Question 27 - Bar graph showing find frequency for a recently-used document.....	176
Figure 84: Box plot showing how satisfaction differs by search success frequency for a recently-used document.....	176
Figure 85: Question 28 - Bar graph showing find method for an old document.....	176
Figure 86: Box plot showing how satisfaction differs by find method for an old document.....	177
Figure 87: Question 29 - Bar graph showing find frequency for an old document.....	177
Figure 88: Question 30 - Bar graph showing how people use the search facility.....	178
Figure 89: Question 31 - Bar graph showing most common search target.....	178
Figure 90: Box plot showing difference in satisfaction varying with target of search.....	179
Figure 91: Bar graph showing the number of participants who indicate they always or often use each search criterion.....	179
Figure 92: Question 33 - Bar graph showing how many people have failed to find a file.....	180
Figure 93: Question 34 - Bar graph showing common reasons for failing to find a file.....	180
Figure 94: Question 35a - Bar graph showing how many participants view the tree.....	181
Figure 95: Question 35b - Bar graph showing how many participants use the tree to navigate.....	181
Figure 96: Question 35c - Bar graph showing how many participants have the address bar visible.....	181
Figure 97: Question 35d - Bar graph showing how many participants use the address bar to navigate.....	182
Figure 98: Question 35e - Bar graph showing how many participants use the back and forward buttons to navigate.....	182
Figure 99: Question 35f - Bar graph showing how many participants use the 'Up One Level' button on the toolbar to navigate.....	182
Figure 100: Question 35g - Bar graph showing how many participants use the keyboard to navigate.....	183
Figure 101: Question 35h - Bar graph showing how many participants open each folder in a new window.....	183
Figure 102: Box plot showing respondents are less satisfied if they open each folder in a new window.....	183
Figure 103: Question 36 - Bar graph showing most common folder view.....	184
Figure 104: Question 37 - Bar graph showing how many participants always or often participants sort by	

each criterion	184
Figure 105: Question 38 - Bar graph showing how many participants have separate files for document versions	185
Figure 106: Question 39 - Bar graph showing how participants distinguish between file versions	185
Figure 107: Question 40 - Bar graph showing how many participants lose track of which file is the most recent version of a document.....	186
Figure 108: Box plot showing people who sometimes lose track of the most recent version of a document are less satisfied overall.....	186
Figure 109: Question 42 - Bar graph showing how many people accidentally have multiple copies of the same document	187
Figure 110: Box plot showing people who report accidentally duplicating documents are less satisfied overall	187
Figure 111: Question 43 - Bar graph showing how accidental duplication happens.....	188
Figure 112: Question 44 - Bar graph showing reasons for deleting a file.....	188
Figure 113: Question 45 - Bar graph showing use of Recycle Bin.....	189
Figure 114: Question 46 - Bar graph showing how often documents are received from the Recycle Bin	189
Figure 115: Question 47 - Bar graph showing the final destination for inactive documents	190
Figure 116: Question 48 - Bar graph showing how often hard drive space is considered in delete decisions.....	190
Figure 117: Question 49 - Bar graph showing self-reported Windows experience	191
Figure 118: Question 50 - Bar graph showing number of computers used	192
Figure 119: Question 51 - Histogram showing years of computer use.....	192
Figure 120: Question 52 - Histogram showing years in current line of work	193
Figure 121: Question 53 - Bar graph showing age distribution	193
Figure 122: Question 54 - Bar graph showing gender	193
Figure 123: Question 55 - Bar graph showing proportion of Academic and General staff.....	194
Figure 124: Question 56 - Bar graph showing proportion of participants in each department	194
Figure 125: Frequency distribution of total number of files in the snapshot	197
Figure 126: Frequency distribution of total number of folders in the snapshot	197
Figure 127: Frequency distribution of the number of top-level locations in each snapshot.....	198
Figure 128: Comparison of actual and reported use of the Desktop	199
Figure 129: Box plot showing high Desktop users are less satisfied than non-Desktop users	200
Figure 130: Allocation of files between Desktop, My Documents and other locations	200
Figure 131: Bar graph showing distribution of maximum folder depth	201
Figure 132: Histogram showing distribution of mean number of subfolders per folder.....	202
Figure 133: Line graph showing how the average number of subfolders in a folder varies with the depth of the folder	202
Figure 134: How the average number of subfolders in a folder varies with the depth of the folder across each of the three collections	203
Figure 135: Histogram showing distribution of branching factor.....	203
Figure 136: Bar graph showing the percentage of folders containing different numbers of subfolders	204
Figure 137: Graph showing the percentage of folders containing different numbers of subfolders for each participant	205
Figure 138: Graph showing the percentage of folders containing zero, one or two subfolders for each participant	205
Figure 139: Histogram showing distribution of mean number of files per folder	206
Figure 140: How the average number of files in a folder varies with the depth of the folder	207
Figure 141: Bar graph showing total number of files stored at each folder depth	207
Figure 142: Histogram showing the distribution of the standard deviation of the number of subfolders ...	208
Figure 143: Histogram showing the distribution of the standard deviation of the number of files.....	209
Figure 144: Histogram showing number of empty folders.....	210
Figure 145: Histogram showing percentage of folders that are empty.....	210
Figure 146: Histogram showing distribution of proportion of file name duplication.....	211
Figure 147: Histogram showing distribution of folder name duplication.....	211
Figure 148: Graph showing relationship between folder name duplication and file name duplication	212
Figure 149: Histogram showing the total number of shortcuts	214
Figure 150: Bar graph showing percentage of files by file type group	214
Figure 151: Bar graph showing number of participants who have files of each file type group	216
Figure 152: Bar graph showing average folder occupancy by file type group.....	216

Figure 153: Line graph showing distribution of file name length	217
Figure 154: Line graph showing distribution of folder name length.....	218
Figure 155: Bar graph showing age distribution of Academic vs. General staff	221
Figure 156: Initial conceptual model created from interview analysis (repeat of Figure 11)	226
Figure 157: Final conceptual model	234
Figure 158: Document management system primary capabilities	235
Figure 159: Possible altered trade-off between filing and finding.....	238
Figure 160: Relationship between strategy and structure.....	240
Figure 161: Summary of guidelines for development of document management user interfaces	241

TABLE OF TABLES

Table 1: Classifications of organising strategies.....	33
Table 2: Comparison of Research Strategy Frameworks	45
Table 3: Field Study Participant Summary	55
Table 4: Data captured by the File System Snapshot program	108
Table 5: Inductively generated codes	124
Table 6: Proportion of Folders Coded with each Code	127
Table 7: Examples of Multiple-Coded Folder Names	129
Table 8: Combinations of Multiple Coded Folder Names	130
Table 9: Triple Coded Folder Names	131
Table 10: Primary Organisational Scheme	132
Table 11: Attitude statements on Survey Question 2	158
Table 12: Factor analysis pattern matrix (Oblimin rotation).....	159
Table 13: Analysis of Variance results for cluster analysis	222
Table 14: Summary of cluster analysis.....	223
Table 15: Summary of document management strategy clusters	236
Table 16: Summary of user interface guidelines and their justifications.....	254
Table 17: Contributions to knowledge made by this research	259

1.1 RESEARCH PROBLEM

It is a very basic human impulse to collect things. Most people spend their lives surrounded by the things we have acquired: in our homes, in our cars and in our offices, and in bags and pockets wherever we go. Going shopping to acquire new things is a major leisure activity in the western world. We usually acquire or retain things because they have some value to us. The value could be that we expect to use the item to make our lives easier, use it to inform us, or simply that we use them to evoke some memory or sentiment.

Many of the things we surround ourselves with were not acquired by choice, such as junk mail or old bus tickets. Often we keep these things around simply because it is easier than going to the effort of disposing of them.

We spend much of our lives managing the stuff that surrounds us. We decide what to acquire and when, where to put it, how to group it with other items, how to find things we need, and what and when to throw away. There are a range of products available to help us with this process: cupboards, drawers, shelves, filing cabinets, folders, paper trays; as well as numerous books telling how to most efficiently use all these items. These things all help us to offload some of the cognitive effort of keeping track of everything.

The personal computer revolution over the past 20 years means that more and more we are being required to manage things in a digital rather than physical form. We acquire documents, email messages, instant messages, music, videos, photos, web bookmarks and more. Because of the rapid

explosion of this technology, research about how we can manage these things has not kept pace with their use.

Search engine algorithms and interfaces are constantly improving, allowing people to find more relevant information faster, while Computer Supported Cooperative Work initiatives allow people to share information better than ever before. Data Mining aims to allow people to find patterns in masses of data, while Visualisation tools help us to visualise these patterns. Knowledge Management initiatives have been implemented with the intention of improving the quality of and access to relevant information across an organisation, and Digital Libraries aim to make vast quantities of information available and searchable to everyone who can access them.

Information, especially digital information, is no longer a scarce resource; information exists in abundance and human time and attention has now become the scarce resource (Simon, 1997). Information overload is now a recognised problem as people struggle to manage the increasing quantities of information they need to deal with on a daily basis (Edmunds & Morris, 2000).

Organisations often try to address the problem of information overload with better decision support and data processing tools, trying to help employees keep up with the flood of data. There are many individuals for whom information processing (especially digital information) is now a significant part of their jobs. Peter Drucker coined the term 'knowledge workers' in 1959 to describe this group of people (Drucker, 1959). Even as they use their visualisation software and data processing tools, there is another overload creeping up on them. A morass of reports, memos, articles, notes, presentations, graphics, contacts, web URLs, emails, tasks and appointments has slowly but surely been accumulating on their computer. While finding information in databases and on the web is becoming easier, finding information located on their own hard drive is becoming more and more difficult.

Many of these people will spend a great deal of their time using software tools to locate, acquire, manage, communicate, process and otherwise interact with this growing plethora of digital information. Because these tasks occupy such a large amount of their time, it is important that these software tools are usable, that is, they are properly designed to effectively support information management activities. Because these activities are so ubiquitous, even small improvements in the usability of the tools could result in a large productivity gain for knowledge workers.

There are numerous different types of digital information that knowledge workers might be engaging with, including web pages, email, documents, images, sound, video, memos, contacts, appointments and tasks. Each of these different types of digital information has its own particular features and requirements. Because of the relative newness of web, email and multimedia technologies, management of these has been the focus of many research efforts. However, the older and more basic task of managing more ordinary documents has gone relatively unstudied.

Most people store their documents in the hierarchical file system provided by their computer's operating system, and manage these documents through a hierarchical file browser (such as Windows Explorer) (Faichney & Gonzalez, 2001). These file browsers were intended to allow a systems administrator to manage files on a computer (at a time when there were generally only a few hundred files). Additionally, when these were developed, computers were not used by the general public, but by highly trained technicians with a thorough understanding of computer technology. The basic paradigm of the tool has not changed in the decades since its introduction, although the user interface to it significantly improved with the widespread introduction of graphical user interfaces in the Macintosh and Windows operating systems. Despite these improvements, the user interfaces of these systems were not designed for modern document management tasks.

The motivation for this study is to improve user interfaces for personal document management. A basic principle of user interface design is that the design of a tool should be thoroughly grounded in an understanding of how the users work, what tasks they perform and how those tasks are carried out. However, with personal document management, very little research has been done into how people are managing their documents and what the requirements are for document management tools. This knowledge gap needs to be addressed before better tools can be developed.

1.2 AIM

The aims of this research are twofold. The first aim is to improve understanding about how people manage their personal digital documents. In particular, there are three aspects to this: (1) understanding the structures people create, (2) understanding the strategies people use to manage those structures, and (3) understanding the problems people encounter and how they solve those problems. The second aim is to use this knowledge to develop guidelines for improving the usability of personal document management software.

1.3 CONTRIBUTIONS

This research makes theoretical, empirical and practical contributions to our understanding of personal document management.

The theoretical contributions arise from the conceptual framework of personal document management developed in **Chapter 6**. This theory provides a comprehensive understanding of the factors affecting personal document management structures and processes and how these interrelate. This theory will be able to guide future research in personal document management.

The empirical contribution stems from the quantitative data gathered about the structure and use of personal document systems presented in **Chapter 5**. This information will be of use both to future researchers as well as to designers of systems for personal document management.

The practical contribution is in the form of detailed design guidelines and user personas developed in **Chapter 6**. The guidelines can be used to inform the design of new personal document management software as well as the redesign of existing software.

1.4 RESEARCH APPROACH

To achieve this aim and make these contributions, **Chapter 2** provides a review of the current theory and knowledge of personal document management practices, and draws from this a framework containing the known requirements and guidelines for the design of personal document management systems. **Chapter 3** summarises the research questions that arise from the current state of the theory and describes and justifies the selection of three research techniques: semi-structured interviews, file system snapshots and a survey. **Chapter 4** describes the interviews, including the interview protocol and administration, how the File System Snapshot technique was developed and implemented, and the results obtained from these techniques. **Chapter 5** details the survey and file system snapshots that were conducted and shows the results they yielded. **Chapter 6** draws on the findings from these three techniques as well as current literature and examines the significance of these findings, followed by the formal conclusions presented in **Chapter 7**.

Chapter 2

LITERATURE REVIEW

The previous chapter explained the importance of investigating the problem of effectively supporting personal document management. This section provides a critical review of the current knowledge, practice and theory in this area.

Section 2.1 defines and describes personal document management, setting out the basic concepts and terminology that will be used throughout this thesis and positioning this research in the context of the field of Human-Computer Interaction.

Section 2.2 reviews empirical studies conducted of personal information management practices, with a particular emphasis on the studies that have involved examinations of document management. The lack of information specifically about document management will be highlighted.

Section 2.3 reviews a number of prototype systems that have been developed to support personal document management, as well as examining the capabilities of a few notable commercial systems.

Section 2.4 discusses theory relevant to personal document management, noting the lack of theory directly relevant to personal document management. This section will include theory from related fields such as psychology, library science and information retrieval.

2.1 TERMS AND DEFINITIONS

This section defines the key concepts that will be used throughout this thesis relating to document management and provides definitions for all the key terms used.

One of the problems with the current state of HCI research into everyday activities is a lack of agreement on common terms and definitions (Whittaker, Terveen, & Nardi, 2000). For this reason, this section will start by reviewing definitions of personal information management (PIM) and from there develop a definition of personal document management as a subset of PIM.

2.1.1.1 Personal Information Management

Personal Information Management has several definitions. Bellotti, Ducheneaut, Howard, Neuwirth and Smith (2002) emphasise categorisation for later retrieval when they define it as *“ordering of information through categorization, placement, or embellishment in a manner that makes it easier to retrieve when it is needed”*. Likewise, Lansdale (1988) defines it as *“the methods and procedures by which we handle, categorise and retrieve information on a day-to-day basis”*, although he notes that retrieval is not the only purpose for which information is handled and categorised.

Barreau (1995) defines a personal information management system as *“an information system developed by, or created for, an individual in a work environment”*, elaborating on the five functions the system must provide: acquisition, organisation, maintenance, retrieval and presentation. The emphasis in this definition is on the system for supporting PIM, rather than the user or the user’s activities.

Boardman (2004) refines this definition into *“the management of personal information as performed by the owning individual”*. He elaborates on management as consisting of acquiring, organising, maintaining and retrieving, and explicitly conceptualises PIM as a user activity rather than a system activity. **Figure 1** below shows this model of Personal Information Management.

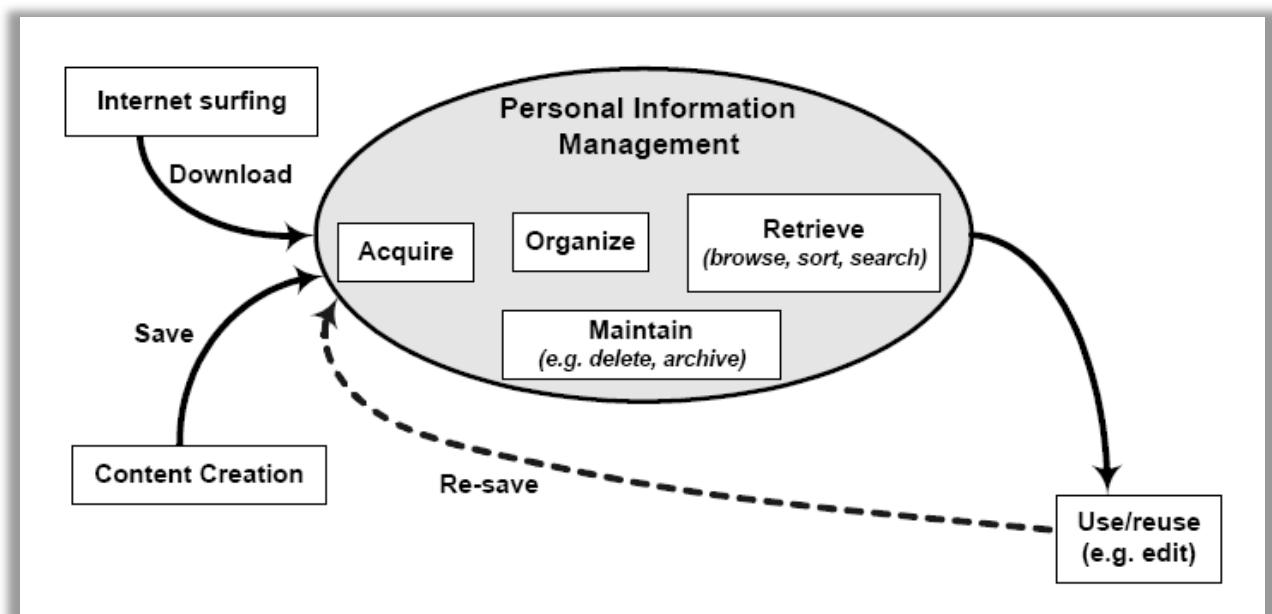


Figure 1: Conceptualisation of Personal Information Management as consisting of acquisition, organisation, retrieval and maintenance (Boardman, 2004, p. 17)

Ownership in this sense implies that the person has control over the document. The owner need not be the author of the document – in fact many people manage documents that they have acquired from others, or work on collaboratively with others. As long as the person has the ability to manipulate their own copy of document (e.g. the ability to rename it, move it or delete it), they are considered to own that representation of that document.

This definition of PIM will be used as the basis for formulating a definition of personal document management in the following section.

2.1.1.2 Personal Document Management

This section will construct a definition of personal document management step by step, through examining and then combining the component terms personal, document and management.

Personal

There are two possible interpretations of the word *personal* in personal information management.

The first interpretation is that personal refers to *information about an individual*. For instance, information an organisation might store about a person, such as their address, date of birth, credit history. This type of information is often personally identifying and may be sensitive. It is usually not controlled or managed by the person the information is about, and hence is often the subject of privacy concerns.

The second interpretation is that it is *information owned by and under the control of the individual*. This is the sense we use when we refer to ‘his documents’ or ‘her email’. The individual has the ability to add to, change or delete this information at will.

This second interpretation is the meaning used in the context of personal information management, and therefore is the meaning that will be used in this research.

Document

Defining a document is surprisingly difficult (Buckland, 1997). A typical dictionary definition (Dictionary.com Unabridged, v1.1) is:

1. A written or printed paper furnishing information or evidence, as a passport, deed, bill of sale, or bill of lading; a legal or official paper.
2. Any written item, as a book, article, or letter, esp. of a factual or informative nature.
3. A computer data file.

These definitions cannot readily translate into a definition of digital documents. The first definition is problematic in that it focuses on paper as the medium, which would exclude digital documents. The third definition is also problematic, since many computer files represent executable programs or program libraries and components and therefore would not be considered by most people to be documents. The second definition focuses on writing of a factual or informative nature. While factual

and informative are debateable (fiction? propaganda?), the main problem with this is in the emphasis on writing. This rules out audio books or images being considered documents, and probably also eliminates spreadsheets and presentations.

The International Telecommunications Union has a definition of a document in a digital context in their Open Document Architecture standard: “a structured amount of information intended for human perception, that may be interchanged as a unit between users and/or systems.” This definition makes no mention of the physical format, meaning that a document can be either physical or digital, and also doesn’t constrain the document to be textual; merely that it is a structured amount of information. It also contains the important idea that a document is a unit that is packaged for a human audience, rather than structured for computer access like a database. Documents can be either static in nature – they are not changed or modified after their initial creation, or dynamic, being continuously or regularly changed or updated over a period of time. This definition includes both types of documents. This definition of document will be used throughout this thesis.

It is important to elaborate on the difference between a document and a file. A document is a logical and human-meaningful package of information. A file in a computer science context is a collection of related data stored as a unit with a single name (The American Heritage® Dictionary of the English Language, 2004). While each document is usually stored in a single file, they are not synonymous. A user may write a report in Microsoft Word which he saves as a file called report.doc. He may then create an updated version of this and save it as a file called report2.doc, and then save a copy of this report in PDF (Portable Document Format) as finalreport.pdf. These are three separate files but they represent a single logical document. Also, a user may split a single logical document such as a long report into multiple smaller files each representing a section of this report. This means that there is not always a direct 1:1 correspondence between a document and a file, particularly with dynamic documents that are modified over time.

Management

Managing refers to directing or controlling the use of something (The American Heritage® Dictionary of the English Language, 2004). In this context, it refers to the activities and actions involved in controlling and using personal information. Following Boardman’s (2004) modification of Barreau’s (1995) management activities, these are the sub-activities that make up management:

Creation and Acquisition. Documents enter a collection either by being created by the user or being acquired from another source. Common sources include receiving documents via email, downloading them from the web and copying them from another computer or memory device.

Retrieval. Documents are retrieved from a collection periodically in order to use them for various purposes, including reading, editing, sending to others, printing or performing organisation and

maintenance activities with them. Document editing, printing and use as attachments in an email system are not part of the scope of document management.

Organisation. Organisation refers to the process of arranging and categorising documents, as well as applying metadata to documents. This includes renaming of documents and folders.

Maintenance. Maintenance activities include deleting documents that are no longer needed, making backup copies of documents, and moving documents that are no longer regularly used to an archive storage location.

While in theory it is possible to ‘manage’ a single document, in practice the subject of the management efforts is a collection of documents. Also, the subject of the document is not important. They may be work related documents, but they equally may be documents related a hobby or private information.

2.1.1.3 Personal Document Management

From the combination of the above definitions, we can reach the following definition of personal document management:

Personal document management is the activity of managing a collection of digital documents performed by the owner of the documents.

The unit of analysis in personal document management is an individual user and the collection of digital documents he or she owns. The process of document management incorporates the creation/acquisition, retrieval, organising and maintenance activities described above, provided they are performed by the document owner. Personal document management is an activity that is performed intermittently, embedded in the daily life of users.

2.1.1.4 Comparison to related terms

It is helpful in coming up with a definition of personal document management to compare it with similar terms and concepts.

Personal Information Management

Personal information management is the parent term of personal document management. The primary difference is in the broader definition of information which encompasses any type of digital information (e.g. emails, tasks, calendar items, contacts) in contrast to the focus on documents as a subtype of information.

Information Management

Information management is “the application of management principles to the acquisition, organisation, control, dissemination and use of information relevant to the effective operation of organizations of all kinds” (Wilson, 2003). It is distinguished from personal document management

through both the emphasis on the organisation rather than the individual, and through the application to all forms of information, structured and unstructured, rather than only documents.

General Information Management

The term General Information Management has been used to describe the type of information management performed by librarians or other professionals to organise and manage information on behalf of others (Bergman, Beyth-Marom, & Nachmias, 2003). Organising and managing information so that it is applicable for a large number of people with differing requirements is a different problem from managing information for an individual's own personal use.

Document Management

The term Document Management can be defined as *"the process of overseeing an enterprise's official business transactions, decision-making records, and transitory documents of importance, which are represented in the format of a document"* (Sutton, 1996). It usually occurs in an enterprise context, overlapping with the term enterprise content management. It differs from personal document management in that the focus is on an organisation, and that usually the content of a knowledge management system has been structured using a taxonomy or ontology by a librarian or information architect (making it a type of General Information Management). The goal is to find a taxonomy that encompasses all of the content across the organisation and that is generic enough that all employees can use it.

Knowledge Management

Knowledge Management is the process of identifying and leveraging the collective knowledge in an organisation in order to improve an organisation's competitiveness (Alavi & Leidner, 2001). It is similar to Document Management in that both are at the level of the organisation. Knowledge management is a broader term and usually includes document management plus measures to capture tacit knowledge, make visible the knowledge resources in an organisation and to reuse various forms of knowledge.

Information Retrieval

The focus of Information Retrieval (IR) is dealing with the representation, storage and access to information items (Baeza-Yates & Ribeiro-Neto, 1999). The term is usually used in the context of locating information in large information sets such as the internet or digital libraries, and the focus has historically been on classification and indexing (Järvelin, 2003). However, IR is an integral part of document management.

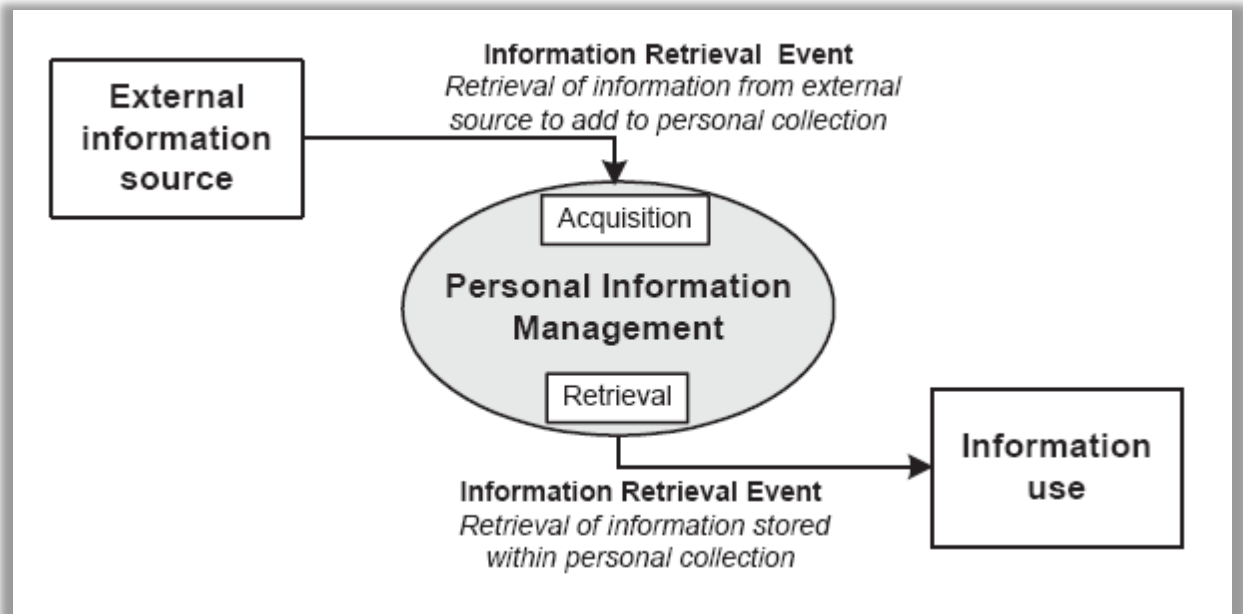


Figure 2: Relationship of Information Retrieval to Personal Information Management (Boardman, 2004)

The tools and techniques of IR are used as part of the process of personal document management in two ways. The first is that many documents become part of an individual's personal documents as a result of a retrieval event – perhaps locating the document on a corporate intranet or digital library. The second is that retrieving a document from the individual's personal document collection is a special case of IR.

Boardman (2004) shows this relationship between Information Retrieval and Personal Information Management in **Figure 2** above.

2.2 EMPIRICAL STUDIES OF PERSONAL INFORMATION MANAGEMENT

This section reviews empirical studies that have been conducted on personal information management. Since so few studies have looked specifically at document management, studies of email organisation, internet bookmark management and paper documents are also included as there are likely to be similarities in behaviour in managing these different types of personal information. These studies examine user behaviour surrounding personal information management and thus generate understanding of user needs and requirements which can inform the development of better tool support.

The process by which researchers can influence user behaviour was described in the Task-Artifact Cycle (Carroll, Kellogg, & Rosson, 1991). Requirements are derived from user needs or problems in creating a task, which motivate the development of an artefact (e.g. computer software) to support it. This artefact can influence the way the task is performed by offering additional possibilities. It is unlikely that any artefact perfectly fulfils a user's requirements, and so there will be a cycle of requirements

gathering based on task problems leading to the development of a new artefact which leads to new task behaviour. The task and the artifacts to support it co-evolve, as shown in **Figure 3** below.

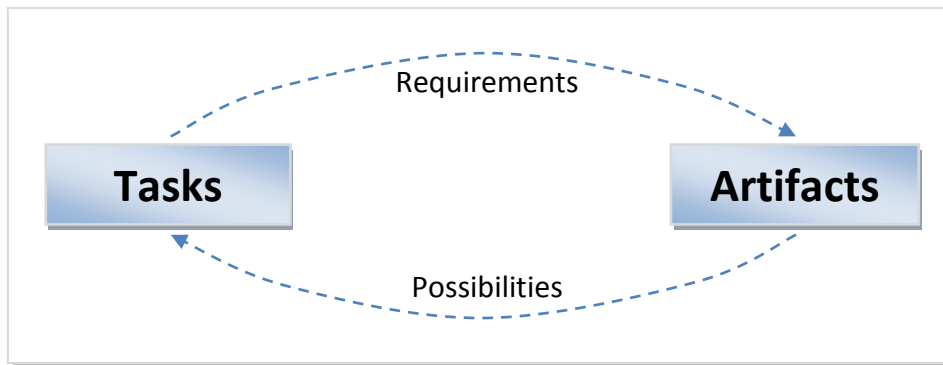


Figure 3: Task-Artifact Cycle (from Carroll et al., 1991)

Different types of research intervene in this cycle in various ways. Empirical studies of user behaviour in performing tasks leads to an improved understanding of requirements for the design of artefacts. Prototypes can be built of alternative artefacts to provide new possibilities to the user, which can then be evaluated in use to assess the difference in task performance. Both types of research support theory building. This theory can be in the form of task models, design guidelines, and artefacts which can be validated by further research.

Empirical work affects the task-artefact cycle by deepening knowledge of user's tasks, and associated problems, performance and requirements. This empirical work usually takes one of two forms: field studies or controlled studies. Field studies are usually conducted in naturalistic contexts and aim to uncover and study real world behaviour as much as possible. Users are often observed conducting real tasks in their native context in order to get the most realistic and relevant information possible. Controlled studies try to collect more objective information in a structured laboratory environment, often constraining the environment and specifying the tasks to be performed. They enable better comparisons to be made, but at the cost of losing ecological validity.

Studies of information behaviour primarily focus on one type of information at a time, with studies looking at paper documents, email, web bookmarks and digital documents. There has also been a study looking at the challenges of managing across collections and how people differ in their information behaviour between the collections. Each of these will be examined in turn in the upcoming sections.

2.2.1 Paper documents

The seminal work in the field of personal information management is Tom Malone's 1983 study titled '*How Do People Organize Their Desks?*' (Malone, 1983). He studied how people used paper files in their offices and identified two distinct strategies: '*neat*' and '*messy*'. In a neat office, the person tried to designate a category for every document and place it the location corresponding to that category. The

location may have been a folder inside a filing cabinet, a paper tray, or a named pile. In the messy office, the person would tend to pile up documents over time, in a less structured way. In both offices, files and piles are the basic building blocks of paper document management. In addition Malone (1983) noted that the organising of information has a purpose beyond enabling the owner to find the information again, it is also an important way of *reminding* them of tasks to be performed. Frohlich and Perry (1994) confirmed this important use of paper for reminding in their study of the use of paper.

Another study of paper files also identified the 'messy' strategy, finding a "*general lack of motivation towards the upkeep of elaborate filing systems - in fact, the less time spent filing the better*" (Cole, 1982, p. 60). He also notes a tendency to have started an elaborate file system, but abandoned it, but at the same time finding that filing is a superior strategy, since if users interact with information often they are more familiar with its structure and location, and don't require as much assistance to locate it. Cole identifies three types of information: action information, personal work files and archived information. He observed that participants often used a different organisation for the different types of information, with action information being organised spatially if at all. Archived information is often formally organised in a categorised way, while personal work files are often partially spatial and partly categorical. He noted a preference for location based retrieval rather than a categorical or semantic one, and in particular noted that time is a poor organising dimension, since "*chronological awareness is not sufficiently developed to serve as a basis for efficient search strategies*" (Cole, 1982, p. 61).

Kwasnik (1989) conducted a study investigating the dimensions people use when they talk about their physical documents in their offices. She found 35 dimensions, which could be grouped into seven broad groups:

Situation Attributes, such as source, use, circumstance, and access; Document Attributes, such as author, topic, and form; Disposition, such as discard, keep, postpone; Order/Scheme, such as group, separate, and arrange; Time, such as continuation, duration, and currency; Value, such as importance, interest, and confidentiality; and Cognitive State, such as "don't know," and "want to remember." (Kwasnik, 1989, p. 208)

More recently, Whittaker and Hirschberg (2001) conducted a qualitative and quantitative study of 50 knowledge workers in a research lab at the time of an office move, in order to test whether filing is indeed superior, and investigate what factors influence choice of paper handling strategies. This allowed them to quantify the amount of paper (in terms of movers boxes), and had forced the participants to recently sort through their paper archives and prioritize them. They found that a piling strategy has many advantages including easier access to recent information and ease of cleaning up their archives, but found that the approach didn't scale well to large document collections.

They observed that people adopting a filing strategy can suffer from '*premature filing*' where people file things of questionable value. They also identify three general difficulties in information processing, regardless of the form of the information:

1. Determining the value of incoming info
2. Deciding whether and how to categorise the data
3. Deciding whether to keep the info in the current workspace or file it away

They further observed that archives are useful in the long term, and that people deliberately keep large archives which they refer to frequently. This contrasts with Kidd (1994) who believes that once the information has done its job of informing, it is no longer needed.

2.2.2 Email

Several studies have attempted to classify styles of email use in a similar way to Malone's 'neat' and 'messy' classifications. One of the earliest was Mackay (1988), who identified '*prioritizers*', '*archivers*' and '*requesters and responders*'. The requesters and responders use email for task delegation; prioritizers concentrate on managing incoming messages while archivers use email to archive information for future use.

Whittaker and Sidner (1996) look more specifically at organising behaviour in email, identifying 'no filers', 'frequent filers' and 'spring cleaners'. The 'no filers' were the email equivalent of pilers, allowing all their email to pile up in the inbox, while the filers attempted to place all their emails into folders. The spring cleaners occupied a middle position between the other two groups, using a 'no-filing' strategy most of the time, but periodically attempting to put their documents into files. Without the folders that others use to aid retrieval, 'no filers' rely on full text search and temporal ordering to retrieve their information. This categorisation was extended by Bälter (1997) to subdivide 'no filers' in to 'folderless cleaners' and 'folderless spring-cleaners' depending on how often they deleted information from their inbox.

In another study, email was identified as a habitat, a location where many users spend a large portion of their daily lives (Ducheneaut & Bellotti, 2001). They found that email is the main form of document exchange, and that some people use email as a document storage repository. More experienced users tended to create more folders in which to organise their email. They also observed users making use of the sorting abilities to assist them in locating emails. Some participants would even use specific folder names (such as beginning with 'a', 'z' or an underscore) in order to force a particular sort ordering.

A more recent study of email behaviour identified two major approaches: 'cleaners' and 'keepers' (Gwizdka, 2004). Cleaners have specific times for dealing with email, and don't keep events or to-do

items in their email. Keepers read email constantly, allowing tasks to be interrupted by email. They keep events and to-do items, and search their email archives.

2.2.3 Web bookmarks

Studies of organising approaches taken with respect to web bookmarks have found similar results to the studies of email, identifying 'no-filer', 'creation-time filer', 'end-of-session filer' and 'sporadic filer', depending on whether and when the user saved web bookmarks during a browsing session (Abrams, Baecker, & Chignell, 1998).

Boardman and Sasse (2004) reported a cross-tool study analysing information behaviour in email, web bookmarks and documents, finding that people tended to neglect their web bookmarks collection, putting more emphasis on organising their email and documents.

2.2.4 Digital Documents

The most significant research to study how people file, find and remind with digital documents was reported by Barreau and Nardi (1995). Their 22 participants were comprised of 4 DOS users, 1 Windows 3.0 user, 16 Macintosh users and one user of OS/2. The study presented these four conclusions:

1. Users preferred to use location based search (browse) for files
2. Users used the placement of files to serve a reminding function
3. Users dealt with three types of information, which were characterised as ephemeral, working and archived.
4. Users did not attach much importance to archiving information.

The study also found differences in the use of subdirectories, but the authors indicate more research is needed in this area.

Location-based search

The conclusion that users prefer location based search was a controversial one and was likened to *"concluding that radio listeners of the 1920s preferred headphones for listening, despite the fact that radios with headphones had not yet been invented"* (Fertig, Freeman, & Gelernter, 1996a, p. 66) .

Although Barreau and Nardi's participants clearly used location based search for the majority of their finding activities, the criticism was that this was because it was the lesser of two evils, not because it is necessarily the best way to retrieve information. Lack of flexibility and scalability of location-based search was singled out as being a problem:

"Location-based storage assumes a small information collection (basically what the user can remember) and does not scale to large collections of information. But information is not always needed in the same way (and thus, not in the same location) as it was originally. Archived information is often needed in a context that is different from the one in which it was created, and in a different location" (Fertig et al., 1996 p.67)

The authors defended their conclusions, pointing out that location based search reduced the amount of information the user needs to remember (Nardi & Barreau, 1997). This trade-off between recall and recognition was identified by Lansdale (1988), who noted that it takes less cognitive effort to browse a directory until the correct name is recognised than to try and remember the precise filename in order to formulate a search. In their original article, they noted that people pay attention to filenames, but the intention is to facilitate recognition of the contents of the file. People do not name files with a view to recalling the exact file name at a later stage.

The authors also attribute the preference for location-based search to the sense of control the user feels during the browsing process as opposed to "*sitting there waiting for the computer to return a list of files that may or may not be relevant*" (Barreau & Nardi, 1995, p. 41).

Reminding

Following on from Malone's (1983) identification of the importance of using document location for reminding on physical desktops, Barreau and Nardi's (1995) participants left files in obvious locations as reminders to do something with them. Fertig et al disagree with this, claiming that despite the availability of time management, and to-do applications, there is little integration of the reminding function into information management solutions (Fertig et al., 1996a). Using location to support reminding provides no guarantee that the reminding will actually occur in time – to some extent, this is left to chance. They see the use of location as a reminding function as a coping strategy adopted only because the software doesn't support anything better.

It is clear from the study that users do attribute meanings to particular locations. For instance, there is no inherent meaning or reminding function in a paper tray. It only is useful to us because we ascribe to it the meaning of 'in-tray' with things to be processed. Likewise, certain locations in the electronic landscape take on the meaning of a reminding location. Rather than ignore this, it would be best to incorporate it, and perhaps augment it with a more robust reminding function, such as the ability to associate a due time with any document.

Types of Information

Barreau and Nardi (1995) update Cole's (1982) classification of types of information with "ephemeral information", "working information" and "archived information" (Barreau & Nardi, 1995). Ephemeral information includes some email, to-do, memos, and other information that must be acted upon in the near future. It generally has a short time horizon measured in hours or days. Working information is related to a user's current projects, and is used frequently during its lifespan. It typically must be kept readily accessible for a moderate amount of time, measured in weeks or months. Archived information is information that may be useful in the future - it is long term storage, with a time horizon of months or years.

Importance of Archiving

Barreau and Nardi (1995) believe that archived information has been overemphasised by researchers because it is important for researchers, but is not necessarily useful for other types of worker. *"Every user in the study indicated that their attempts to establish elaborate filing schemes for archived information failed because they proved to require more time and effort than the information was worth"* (Barreau & Nardi, 1995, p. 42). The authors suggest that the primary use of information systems is to provide cues and a space people can use to remind and organise their ideas, with the use of computers as a repository of stored information being a secondary use. In this they agree with Kidd (1994), who claims that archived information is of very little relevance to knowledge workers. Kidd subscribes to what Whittaker and Sidner (1996) call a 'one touch' model – the idea that the knowledge worker reads the information, internalises it into their own knowledge space, and then discards it as no longer needed.

Interestingly they find that *"some users reported they would recreate a memorandum, paper, or study rather than store, locate and edit an older version for a new purpose"* (Nardi & Barreau, 1997, p. 3). This suggests that people are so loathe to store material that they will re-create work, which is contradicted by other studies (particularly Frohlich & Perry, 1994; Whittaker & Hirschberg, 2001) which found that people are so apt to store information that half of what they kept was readily available elsewhere (Whittaker & Hirschberg, 2001). This was found to be true of non-researchers as well as researchers.

It must be noted that at the time of these studies, the majority of the participants did not have hard disk drives, and any archiving that was done used floppy disks. While that may require so much effort that the participants abandoned the task, this does not necessarily mean that the information is not valuable. Rather, if the current technologies made archiving impossible or prohibitively difficult, this points to the need for more work to be done to improve document management user interfaces in order to facilitate both archiving and the retrieval of information from archives.

One study of document narratives updated Kwasnik's (1989) research to describe the stories people tell about their digital documents. The author found that the dimensions most commonly used in narratives about documents were: Time, Place, Co-Author, Purpose, Subject, Other Documents, Format, Exchanges, Tasks, Storage and Contents (Gonçalves & Jorge, 2004 p.248).

The only other more recent study to look at digital documents was recently conducted by Richard Boardman (2004). He analysed information behaviour across three collections: documents, email and web bookmarks with the intention of analysing difficulties people had in managing their information collections across tools. He found that people could be categorised as either 'pro-organising' or 'organising neutral', but that people didn't always adopt the same strategy across all collections. People

were more likely to be 'pro-organising' in their document collection and email than they were in their web bookmarks.

2.3 PROTOTYPES AND SYSTEMS

Much research on information workspaces has attempted to create new workspace software that is different from the current predominant hierarchical system. Proposed systems include those with a primarily temporal paradigm such as Lifestreams (Freeman & Fertig, 1995), primarily spatial paradigm such as Data Mountain (Cockburn & McKenzie, 2001; Robertson et al., 1998) and a primarily logical paradigm such as Haystack (Karger & Quan, 2004), Presto/Vista (Dourish, Edwards, LaMarca, & Salisbury, 1999a) and the Semantic File System (Gifford, Jouvelot, Sheldon, & O'Toole, 1991). In addition, there have been some combinations such as TimeScape, a combination of temporal and spatial information management (Rekimoto, 1999b). Other systems have tried to provide a logical interface to paper documents (Bellotti & Smith, 2000; Rao, Card, Johnson, Klotz, & Trigg, 1994; Trigg, Blomberg, & Suchman, 1999), and others have explored using artificial intelligence for automated classification (Crawford, Kay, & McCreath, 2002; Mock, 2001; Segal & Kephart, 1999).

Arguments for each of these organisational paradigms can be made on psychological grounds. It is known that people make heavy use of spatial cues for organising information (Cole, 1982; Lansdale, 1988; Malone, 1983). Locations are often associated with particular information, and people are frequently able to remember where they put something. Temporal information can provide useful cues for organising information, particularly to aid recognition, however in most people 'chronological awareness' is not sufficiently developed to serve as a basis for efficient search strategies (Lansdale, 1988 p.61). Logical information structuring is based on keyword or content assignments. While it takes advantage of our classification ability and knowledge of topics, the major disadvantage is that it cannot utilise the visual cues at which the human perceptual system excels.

Another issue with many of these proposed systems (such as Placeless Documents), is that they require the user to supply metadata about the document. As encountered in many knowledge management initiatives it is difficult to get users to enter metadata about their documents (Kao, Quach, Poteet, & Woods, 2003). Users are busy getting on with their work, and aren't really concerned about managing their documents beyond the minimum required to ensure the document doesn't disappear into oblivion.

At the other end of the spectrum are commercial systems like the Google Desktop (desktop.google.com), and Copernic Desktop Search (www.copernic.com), which theoretically make organising dimensions unnecessary, since they can locate documents by means of full text searching. Although the full text searching is attractive, it is not a full replacement for organising document, since

browsing through an organised collection of documents gives you an overview of what is available, as well as the ability to see how different items are related to each other. This doesn't happen with a full text search system, which can only retrieve documents matching a query you formulate explicitly. Sometimes users don't need to retrieve a specific document, but just to "*see what information I've got related to X*" (Svenonius, 2000).

Despite the number of different paradigm implementations created by researchers, there is no clear evidence regarding which is better. The evaluation of many of these prototypes has been informal, with the researcher and their colleagues using the system and feedback is only anecdotal.

The following sections first describes the history of the development of personal document management systems, and then presents a more detailed discussion of the document management functions of Windows XP, which is currently the most widely used system for personal document management (NetApplications Ltd, 2008). This is followed by brief overviews of the most influential workspaces proposed by researchers as well as some of the most widely used commercial personal document management systems.

2.3.1 Personal Document Management System History

The origins of personal information management are usually credited to Vannevar Bush and his *Memex* system (Bush, 1945). He describes a system where a user could archive all the information they ever came across on microfiche, and navigate through the information in order to create links between various parts of it. The desk he envisaged was a place where the user could go to access all their personal information as well as any reference material they found useful to store.

The first computer systems on which it was possible to store personal files were multi-user time sharing computes developed in the 1960s. *Multics* was the first of these to allow files to be arranged in a hierarchy, with each user given a directory in which they could create subdirectories and files (Corbató & Vyssotsky, 1965). It was accessed through a command line terminal interface that bears some similarities to modern Unix and Linux systems. Unix is a descendent of Multics, and provided similar command line access features to a home directory.

Work on personal computers was also proceeding during the 1960s, with Douglas Engelbart's desire to realise Vannevar Bush's ideas of using technology to augment human activity (Engelbart, 1962). His team were pioneers in developing the mouse and a graphical user interface with multiple windows. Further technological progress resulted in Xerox developing the first personal computers, beginning with the Xerox Alto in 1973 (Xerox, 2006). This was the first direct manipulation graphical user interface, contrasting with the command line access available on the UNIX platform. Document management was

provided through a directory manager interface called Neptune in which files could be moved and deleted with the mouse.

The first commercial personal computer was the Xerox Star in 1981 (Xerox, 2006), which was the foundation of the desktop metaphor which persists to this day. Users were able to arrange icons spatially on the desktop surface, and the metaphor of filing cabinets, drawers, hangers and folders provided a hierarchical directory structure into which documents could be stored. Although the Xerox Star was a commercial failure, it provided the inspiration for other personal computer user interfaces, including the Apple Lisa and subsequently the Apple Macintosh. The Apple Lisa adopted the idea of a spatial desktop and a hierarchical system of folders based on the Xerox Star desktop metaphor. These basic ideas survive relatively unchanged in today's personal computer user interfaces.

All modern personal computer operating systems have a hierarchical file system, in which files can be stored inside directories (folders). This file system contains all the system files needed to run the operating system, as well as all the files need to run the installed applications. Typically, one branch of the hierarchy is set aside for users to store their own personal documents, although in theory users can store their personal documents anywhere in the file system. There are a range of different formats in which files are stored depending on the application used to create the file. Files of different formats are managed uniformly in the file system (largely including executable and system files).

The file system usually places some restrictions on the names of files and folders, such as the length of the name and on the characters it contains. It is common in many systems to have a file extension, typically a period followed by three characters at the end of the file which denotes the type of information contained in the file. This is typically used to associate applications with particular types of files. The file system provides the ability to create and delete files and folders, and move or copy files from one folder to another.

Each file can only exist in one location in the folder hierarchy. When a file is copied to another location, it creates an entirely separate file which has no relationship back to the source file. Most file systems provide the ability to create links or shortcuts to a file, which makes it appear as though the file exists in multiple locations.

2.3.2 Windows Explorer (in Windows XP)

Microsoft Windows Explorer is a file management utility that is available on all versions of the Microsoft Windows operating system. It provides a view of all the files on a computer, and gives the user the ability to structure them into folders which can be hierarchically arranged. The remainder of this section talks specifically about the version of Windows Explorer that is part of the Windows XP

operating system. Windows XP currently has the largest market share of all personal computer operating systems, as **Figure 4** shows.

The Windows Explorer interface provides two panes (as shown in **Figure 5** below). The left pane shows physical and logical drives and folders only, in a tree structure. This can be expanded or collapsed to show different amounts of the hierarchy. The right pane shows the folders and files inside the currently selected folder. Files can be moved or copied between folders by dragging and dropping, or by cutting and pasting. Each file can only be located in one folder at any time, and copies of a file are independent of the original.

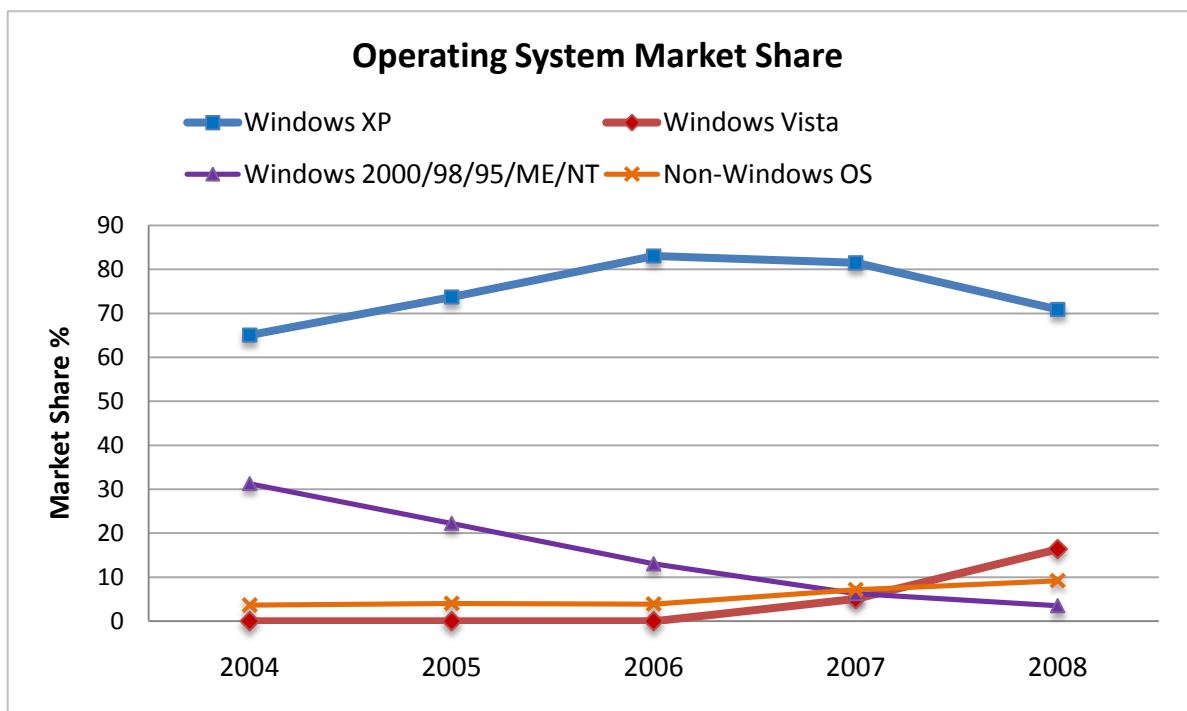


Figure 4: Personal computer operating system market share (NetApplications Ltd, 2008)

There is also an ability to create 'shortcuts' or aliases to a file. Windows Explorer also includes a search feature that allows documents to be searched for by file name or partial contents. There is also an advanced search option, which allows searching by date, file type and file size.

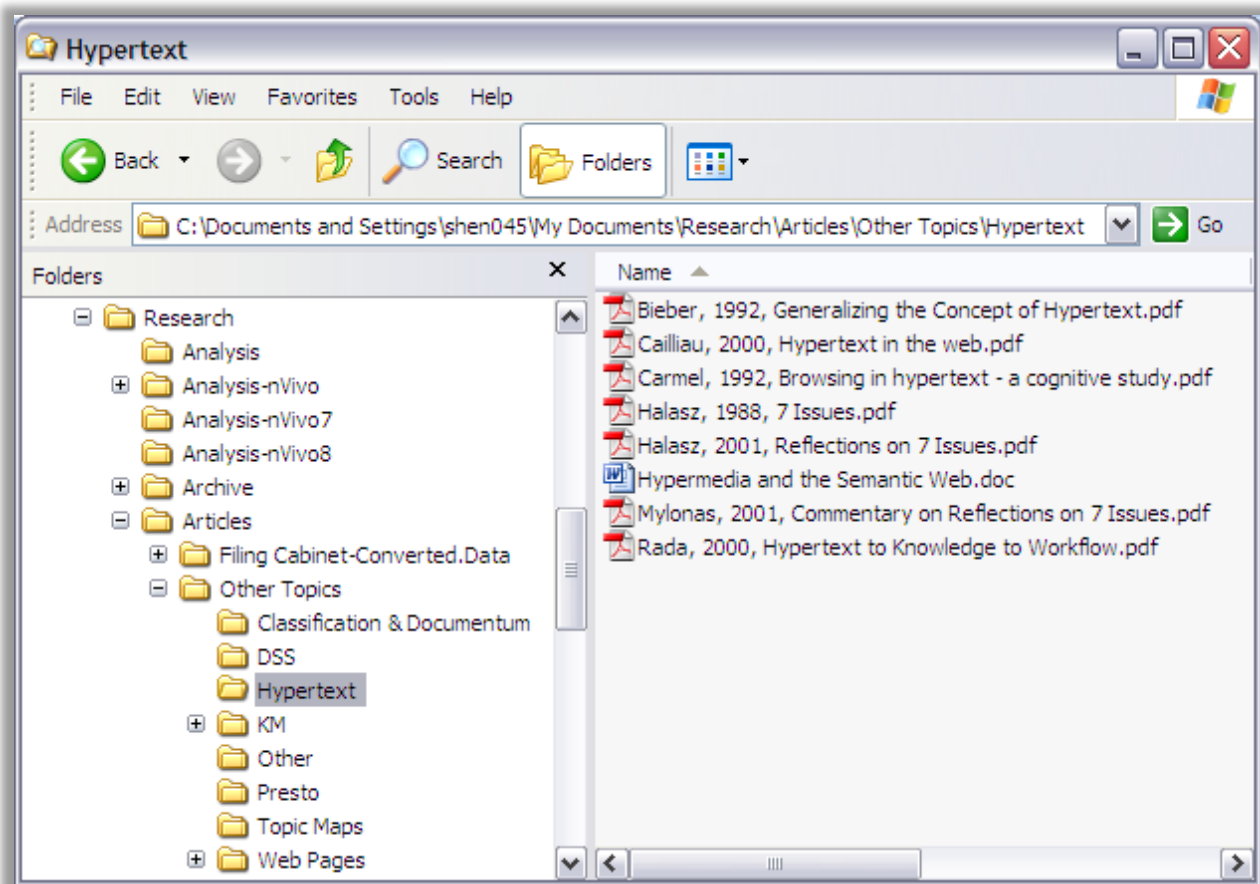


Figure 5: Microsoft Windows Explorer interface (Windows XP version)

Within a folder, files can be viewed with details (**Figure 6**), in a list (**Figure 7**), or as icons (**Figure 8**). Some folders can show customised views depending on the anticipated content. For instance, the 'My Pictures' folder has the ability to view image thumbnails or as a slideshow.

The default details view in Windows XP provides four columns: Name, Size, Type and Date Modified. This is customisable, with an additional 34 attributes available for display. Many of these are particular to certain file types (such as Artist and Album for music files, Camera Model and Date Picture Taken for photos), but a few are applicable to all file types (e.g. Date Created, Date Accessed, Author, Comments).

Name	Size	Type	Date Modified
Lecture 01 - Introduction to the course.ppt	27 KB	Microsoft Office PowerPoint 97-2003 Presentation	2/12/2008 9:24 p.m.
Lecture 02 - Intro to Web Development.ppt	27 KB	Microsoft Office PowerPoint 97-2003 Presentation	2/12/2008 9:25 p.m.
Class List.xls	9 KB	Microsoft Office Excel 97-2003 Worksheet	2/12/2008 9:25 p.m.
Assignment 1.doc	37 KB	Microsoft Office Word 97 - 2003 Document	2/12/2008 9:35 p.m.
Assignment 2.doc	212 KB	Microsoft Office Word 97 - 2003 Document	2/12/2008 9:34 p.m.
Students with test clashes.txt	1 KB	Text Document	2/12/2008 9:36 p.m.
Lab timetable.xls	107 KB	Microsoft Office Excel 97-2003 Worksheet	2/12/2008 9:36 p.m.
Course Outline.pdf	869 KB	Adobe Acrobat 7.0 Document	2/12/2008 9:36 p.m.
Course Schedule.pdf	124 KB	Adobe Acrobat 7.0 Document	2/12/2008 9:37 p.m.

Figure 6: Files shown using Windows XP Details view

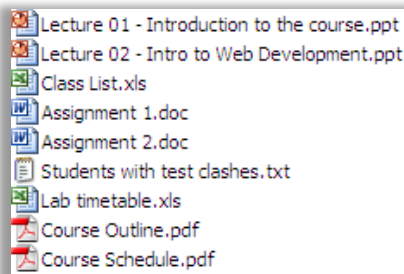


Figure 7: Files shown using Windows XP List view

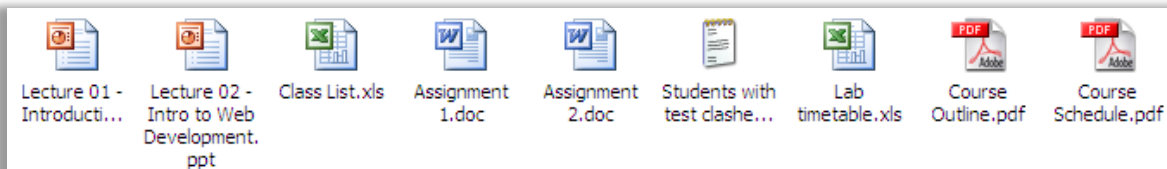


Figure 8: Files shown using Windows XP Icons view

My Documents and the Desktop are special folders created by Windows XP as subfolders of each user's account folder. This means that each user of the computer has their own Desktop and My Documents folder, independent of other users of the computer. My Documents is intended by Microsoft to be the location where most users store their documents, and the Desktop is a place to store shortcuts to frequently used applications, as well as being able to store documents. The My Documents folder is normally located at C:\Documents and Settings\\My Documents, and the Desktop is stored in C:\Documents and Settings\\Desktop, however, Windows XP usually hides these full paths in most views, and automatically creates shortcuts to My Documents on the Desktop and on the Quick Launch shortcut menu to provide easy access. My Desktop is the default location for saving files in the Microsoft Office applications. Only the Desktop rivals it as a convenient storage location and users need to take extra steps in order to use another location for their primary storage.

2.3.3 'Pile' User Interface

One of the earliest prototypes adopted a 'pile' metaphor, based on Malone's (1983) observations of the importance of piles as well as files for organising. The authors noted that unlike physical folders, physical piles let you see at a glance what was in the pile. In addition, piles are easier to browse, allowing the edges to be scanned, the pile to be restacked, hinged or spread out (Mander, Salomon, & Wong, 1992). Their system also tried to introduce automation using the metaphor of an assistant who took care of routine tasks such as filtering information and filing. The assistants often helped with the filing and organising, but "*typically the assistant would suggest categories and discuss these with the worker before actually filing the material*" (Mander et al., 1992, p. 692), thus acting as an intelligent agent.

The researchers created a Macromedia Director mock up of how a pile interface could work, using a 'messy' pile for user created piles and a neat pile for system created piles. The user can create and edit "scripts" to determine what is in the pile, meaning their piles can act as a dynamic query result set.

A document is added to the pile by dropping it on the pile. The pile can be 'spread out' showing all the documents within it. Also, they use a 'viewing cone', which contains a thumbnail of the first page of each document for the user to browse through. The cursor keys can be used to move through the other pages.

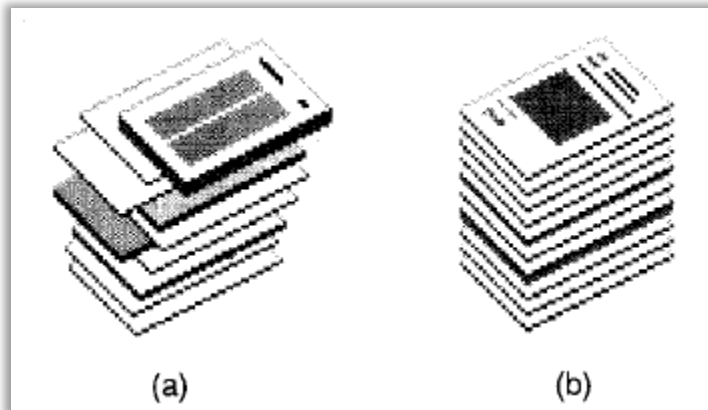


Figure 9: Pile interface showing (a) user-created pile and (b) system-created pile (Mander et al., 1992)

The authors report that their pile interface was generally well received by the participants they tested it with, however it has been criticised. Treglown (2000) uses Lakoff and Johnson's (1987) theory of metaphor to argue that the pile interface embodies the same metaphorical primitives as ordinary folders (Treglown, 2000). Both piles and folders embody the containment schemata, based on the idea of items being contained within another item.

Prior to the development of this interface, Lansdale (1988) was careful to warn that it is not necessarily a good idea to try and directly translate physical mechanisms into electronic systems, since many of the devices (such as unstructured piles) are not representative of anything intrinsically useful, and do not cater for any particular user need, but are a way of coping with current inadequate systems. He specifically singles out the idea of developing an interface based on piles as being equivalent to the idea of "*building planes that flapped their wings*" (Lansdale, 1988, p. 56).

However, in the physical office, Malone (1983) noted a difference between merely using piles as an organising primitive along with files, and being a 'messy' piler. Used by a 'neat' person, the piles had the similar semantics to folders – they grouped together related documents in a physical location – but with the additional benefit of higher visibility. In contrast, the piles created by the 'messy' person tended not to have any coherency or meaning. Mander, Saloman and Wong's (1992) piles interface supports only the former use of piles, providing a visually richer interface than a standard folder, while retaining the same containment semantics.

2.3.4 Lifestreams

Lifestreams was developed by Fertig, Freeman and Gelernter as a way of organising all the electronic documents associated with a person during their lifetime (Fertig, Freeman, & Gelernter, 1996b). It takes a primarily temporal approach, with all documents being represented as a one-dimensional stream through time. The interface is shown in **Figure 10** below. Although the temporal metaphor is very powerful, there are some difficulties that arise when it is the primary mechanism for organisation.

For instance, in order to retrieve a previously worked on document, you need to know when you created it. Another problem arises with which date associated with a document should be used to determine its place in the stream. If creation date alone is used then in order to access a document that was worked on every day for year, the user would have to go back to the creation date to open it every time. Conversely, if it were sorted by last access date, inadvertently opening a document would mean it would change its place in the stream, making it more difficult to locate in future. A combination of both would lead to multiple instances of the same document appearing in the stream, potentially confusing things.

One powerful aspect of the Lifestreams model is the ability to create a document with a future date. This in effect creates a reminder, since the reminder won't appear until the date has passed. This is a useful effect for a reminder, but makes it difficult to see upcoming reminders. No user studies have been performed on this system. Lifestreams also allows documents to be logically grouped into sub-streams.

Another criticism of the Lifestreams metaphor is that it does not use the most common temporal metaphor (Treglown, 2000). The standard temporal metaphor is the 'moving time' metaphor, which has a stationary observer with the future in front and the past behind. By contrast, Lifestreams has the past stretching out in front of the user, and so their design conflicts with the everyday perceptions of most people.

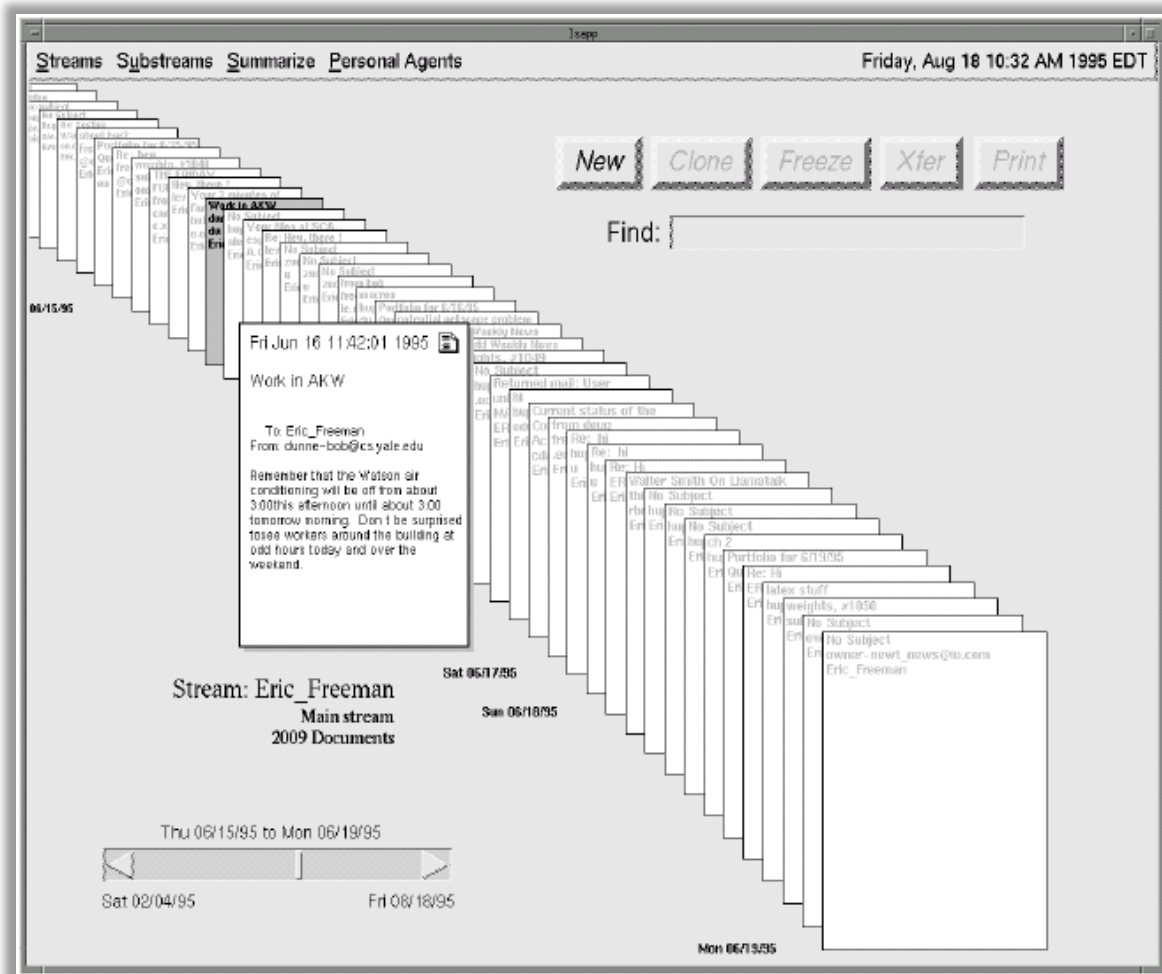


Figure 10: Lifestreams interface (Fertig et al., 1996b)

2.3.5 TimeScope

TimeScope provides a combination of a spatial and temporal organisation of documents (Rekimoto, 1999a). Documents are laid out spatially on a desktop, which is the only workspace available – there are no folders or other grouping mechanisms. When a document is no longer needed, it is deleted from the desktop. To retrieve previously worked on documents (after they have been deleted from the desktop), the user must ‘travel through time’ back to the time when the document was on the desktop. Thus the system allows navigation through previous desktop states.

Figure 11 shows the interface of the TimeScope application. The bar down the left shows the current position in time. In the top right is a time travel dial, allowing desktops from other points in time to be navigated to.

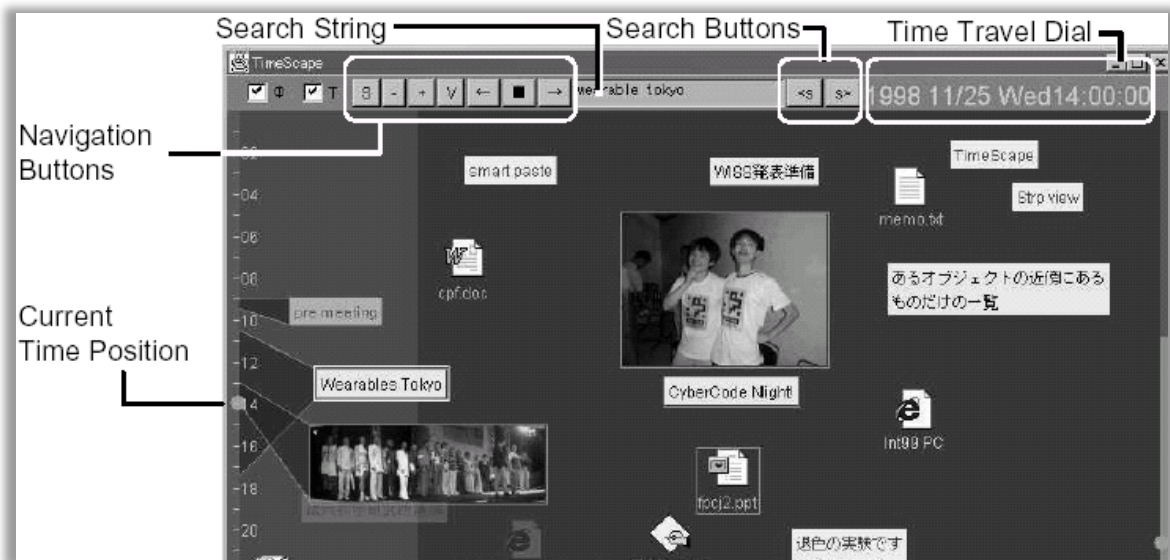


Figure 11: TimeScape interface showing static desktop (Rekimoto, 1999a)

Figure 12 below shows the timeline view of the desktop. It shows the current desktop contents in oblique view, and has bars showing the presence of these documents before and after the current time. Other documents that are on the desktop in either the past or the future can also be seen in this description.

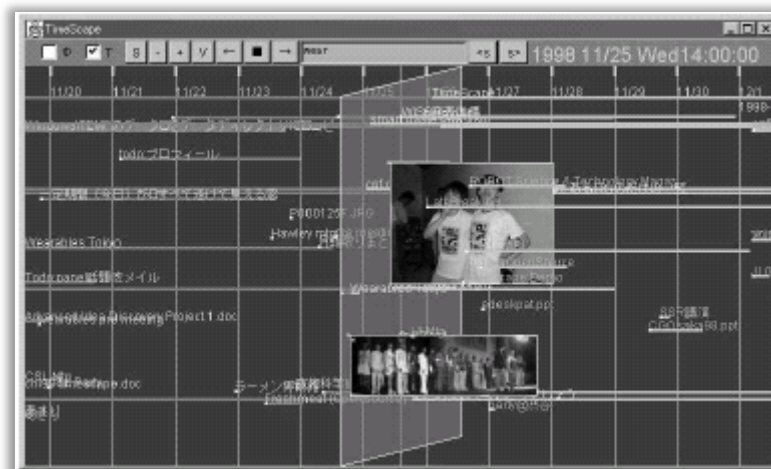


Figure 12: TimeScape interface showing timelines (Rekimoto, 1999a)

This prototype has some of the strengths of the Lifestreams system, but a number of other weaknesses as well. The major problem is that the dominant metaphor is travelling in time, which is not something most people have much familiarity with. Additionally, to remove something from the current workspace requires it to be deleted, which is something many users are reluctant to do, because of the perceived permanence of such an action.

TimeScape supports reminders in a similar way to Lifestreams, creating documents in the future that will appear on the desktop when the current time reaches the time the document was created.

However, it is not possible to create a document and set it to appear in the future – the user must navigate to the future, create the document, and then navigate back to their current workspace.

2.3.6 Presto

Research has been conducted at Xerox PARC into logical document collections, under the umbrella of ‘Placeless documents’ (Dourish, Edwards, LaMarca, & Salisbury, 1999b). This is based on the premise that documents should not be required to have a single defined location. Instead, there exists a large pool of documents, and any subset can be pulled out of the pool as needed according to the document properties. Properties are understood to be any information about the document, including information extracted from it automatically, or metadata manually added to the document.

Presto is a prototype system that manages and stores the properties of documents. Vista is a user interface for the Presto System. The Vista interface is shown in **Figure 13** below. It shows a number of documents on a desktop like interface, as well as a number of groups, both open (large ovals) and closed (piles on left). The triangle icons signify properties, which can be added to the groups, and act as a filter condition.

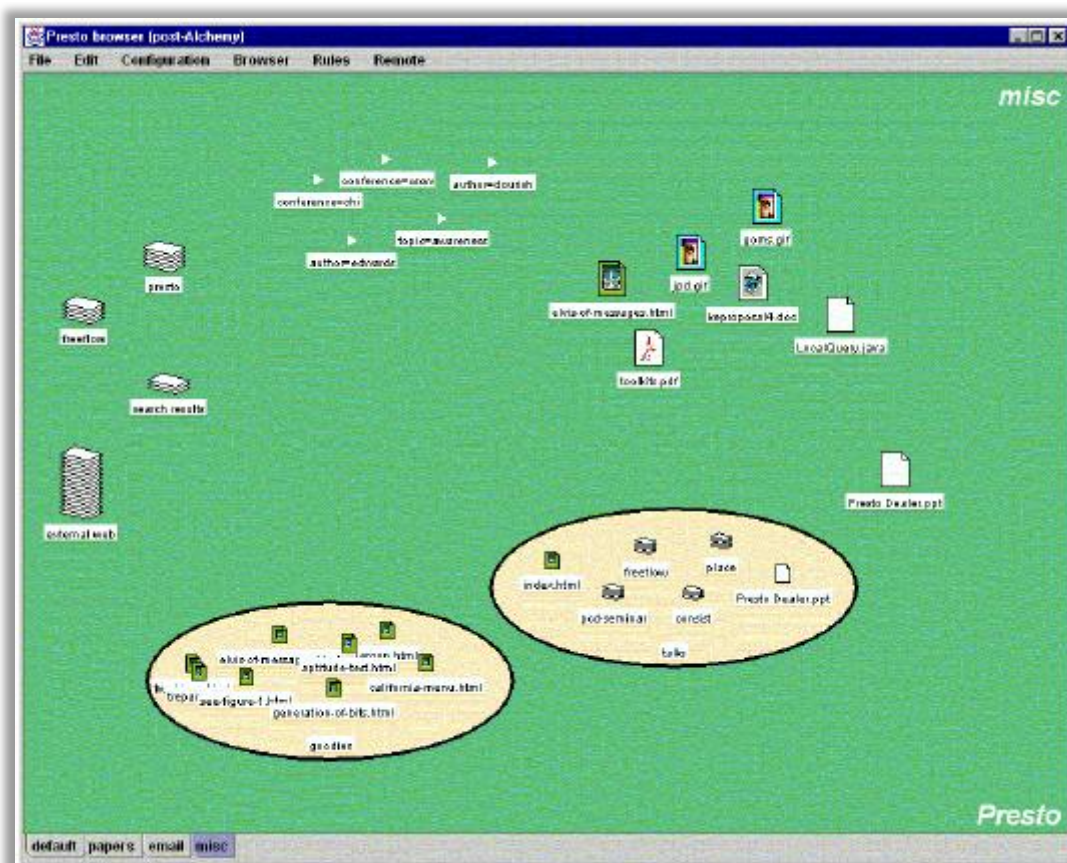


Figure 13: The Vista Interface for the Presto system (Dourish et al., 1999a)

This system has a number of strengths. It doesn't require documents to be named, or to be given a specific location. Documents can always be found through logical properties etc.

However, given people's familiarity with physical objects and spatial metaphors, it might be difficult for people to adjust to a system in which their documents are essentially 'nowhere', but are just in the system.

2.3.7 Haystack

Haystack is a general purpose information management environment designed to improve end-users' ability to store, examine, manipulate, and find their information (Karger, Bakshi, Huynh, Quan, & Sinha, 2003). It is designed to be fully flexible, allowing users to define the types of information important to them, and to define the properties they wish to store about any piece of information. It doesn't have a fixed user interface, but instead allows the end user (or a developer) to specify the UI through editable view prescriptions.

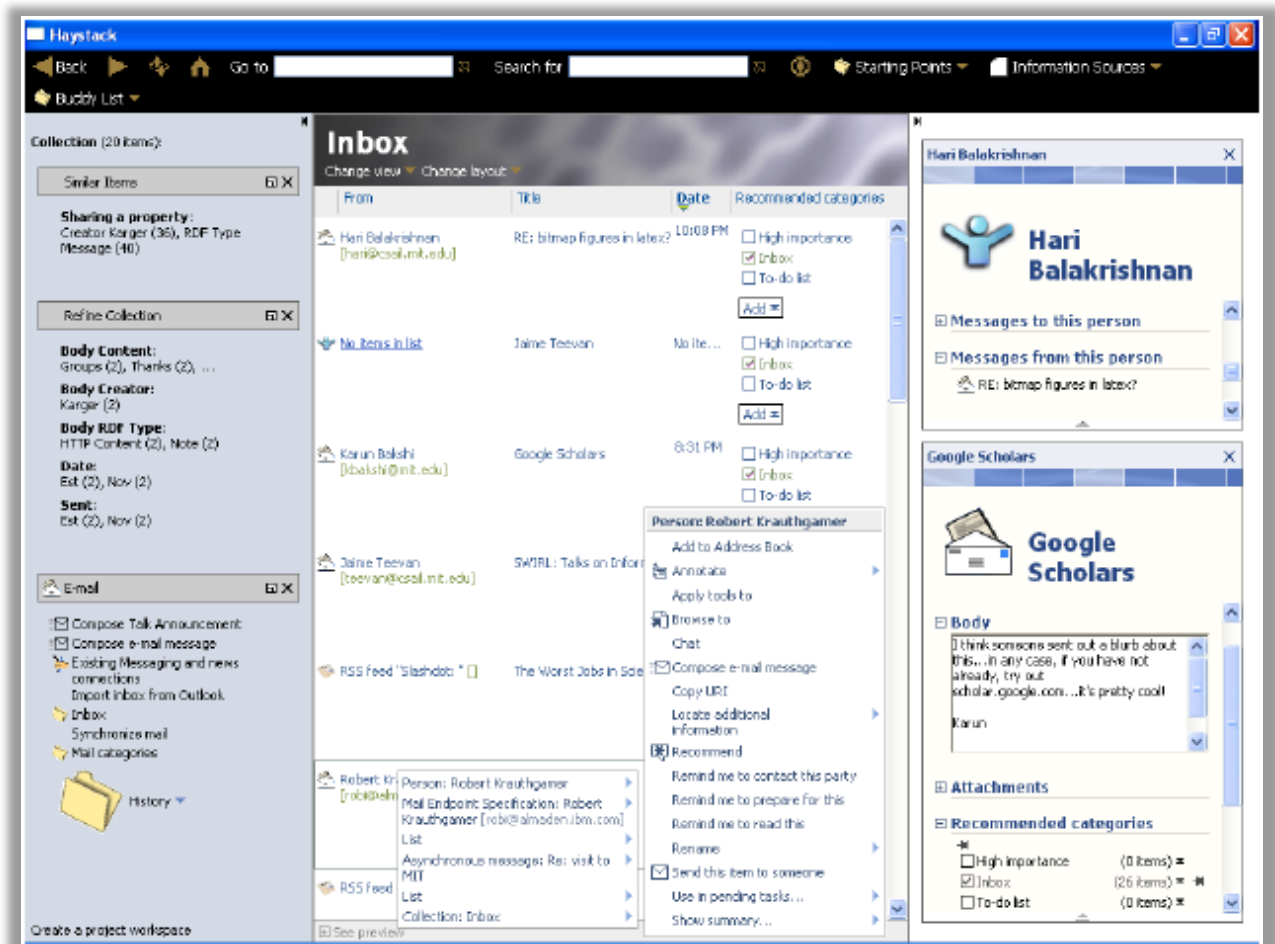


Figure 14: Haystack user interface (Karger et al., 2003)

Haystack is capable of managing not just files but any item of information. It achieves this by assigning a uniform resource identifier (URI) to name anything of interest—a document, an email, a person, or a web page. The named object can then be grouped with other objects, viewed, retrieved

and have arbitrary actions performed against it. **Figure 14** shows a customised Haystack user interface designed to display email items in a way that mimics the capabilities of an email client.

2.3.8 Stuff I've Seen

The "Stuff I've Seen" (SIS) project was developed by researchers at Microsoft in order to facilitate re-finding information that has already been encountered. The system indexes all documents, emails and web pages that a person sees and facilitates searching through the collection. SIS consists of five modules: Gatherer (to access sources), Filter (to decode sources into a character stream), Tokenizer (breaks into words, may also normalize dates, stemming), Indexer, Retriever.

Search results are presented in a list view, with a number of panels that enable filtering by various attributes (as shown in **Figure 15**). The system is primarily a unified information retrieval interface, offering no ability to organise information or to add custom metadata on which to base searches or filters.

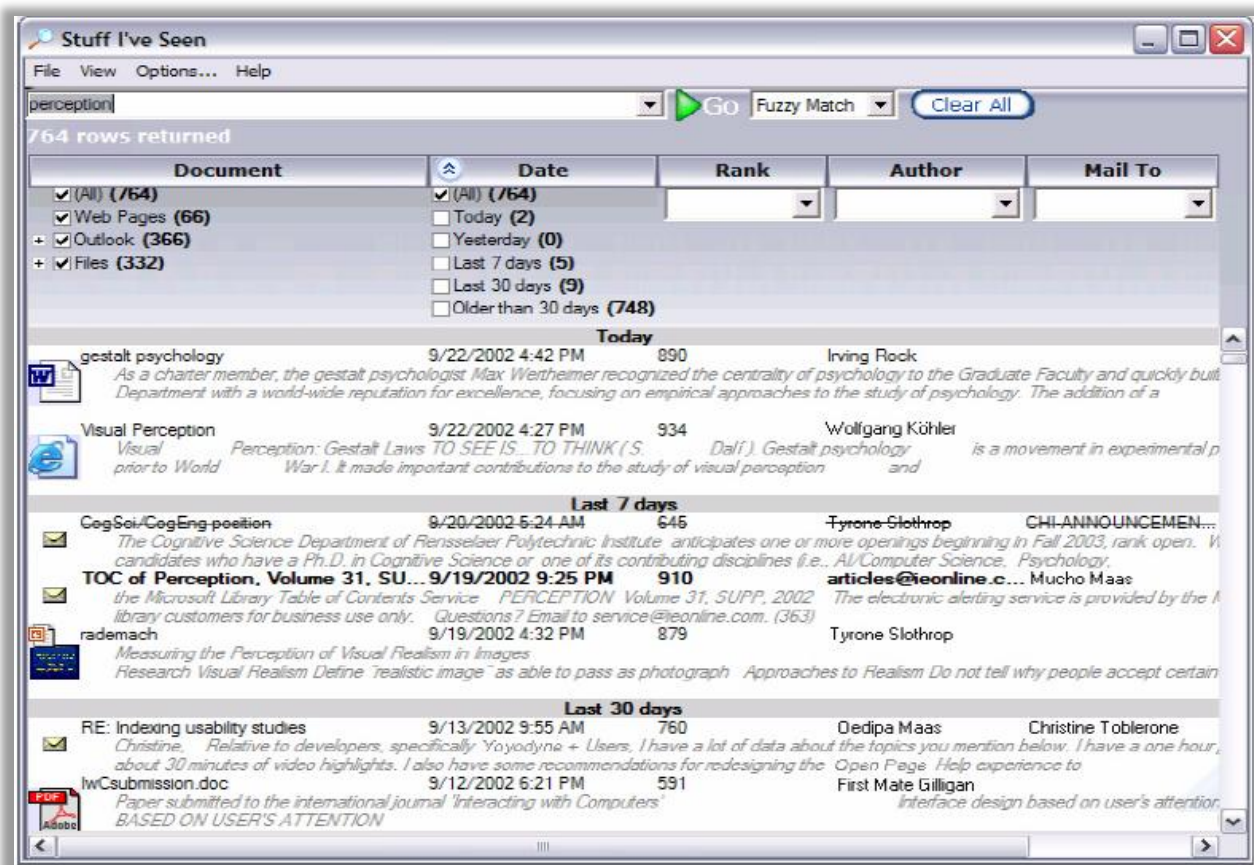


Figure 15: Stuff I've Seen 'Top View' user interface (Dumais et al., 2003)

The system was used by more than 230 Microsoft employees, who found time and people to be the most important retrieval cues. They noted that people frequently search by names, "Their importance is highlighted by the fact that 25% of the queries involved people's names suggesting that people are a

powerful memory cue for personal content. In contrast, general informational queries are less prevalent." (Dumais et al., 2003)

2.3.9 Google Desktop

Google Desktop was launched in October 2004, offering the ability to use Google's indexing technology to index local files, email and web history (Google Inc, 2004). The search is conducted through the browser, with the search interface nearly identical to Google web search, as shown in **Figure 16**. The key strengths of Google's offering are that its continual indexing means that documents and email are available for search almost immediately, and that the search itself is extremely fast.

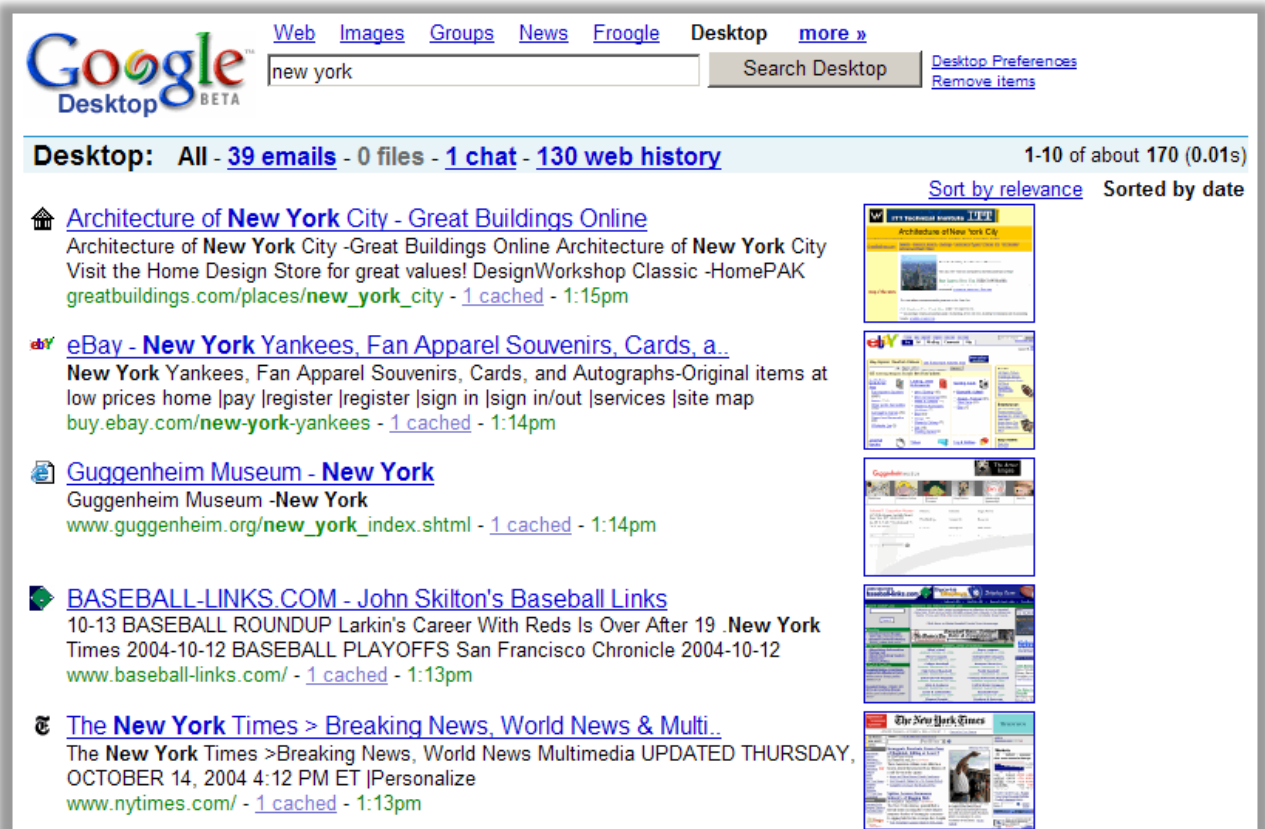


Figure 16: Google Desktop search results

2.3.10 Copernic Desktop Search

Copernic Desktop Search (CDS) is an indexing and search utility for the Windows platform, released free to the public in August 2004 (Price, 2004). CDS builds an index containing the keywords and metadata of documents and email, and then searches this index. After the index is initially built, newly created documents and newly received emails are added within a few minutes. Due to the indexing, the search itself is fast, and offers the ability to easily refine searches based on metadata, in addition to providing a preview of items appearing in the search results (as shown in **Figure 17** below). As well as email, Copernic Desktop Search version 1 handles eight textual file types, plus two image file types and three music file types.

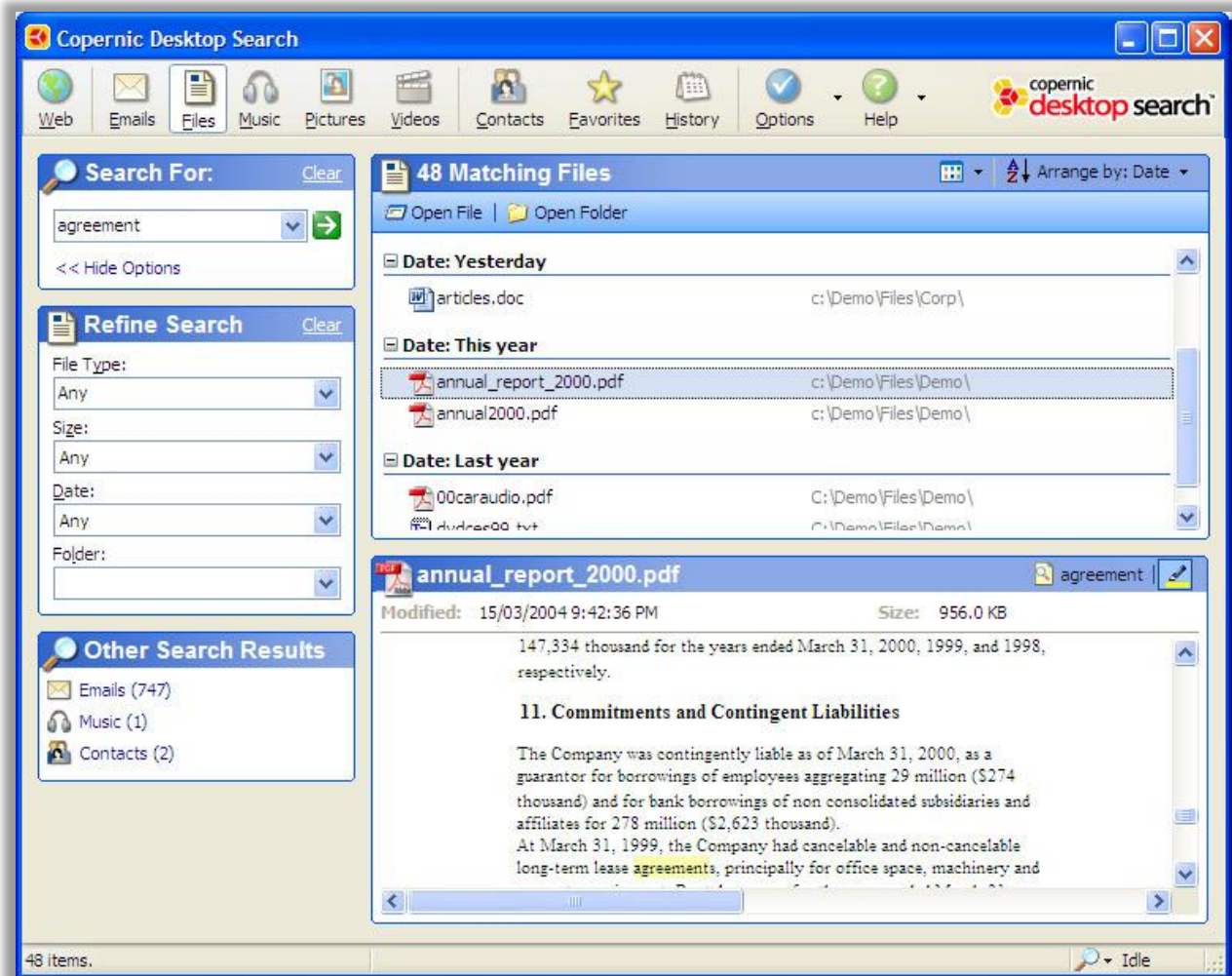


Figure 17: Copernic Desktop Search interface

2.4 THEORY RELATED TO PERSONAL DOCUMENT MANAGEMENT

This section reviews the theoretical findings related to personal document management. There are five main kinds of theory in the field of Human-Computer Interaction (Rogers, 2004, pp. 30-31):

- Descriptive - in the sense of providing concepts, clarifying terminology and guiding further inquiry
- Explanatory – in the sense of explicating relationships and processes
- Predictive – enabling predictions to be made about user performance
- Prescriptive – providing guidance for design
- Generative – in the sense of enabling practitioners to create, invent or discover something new.

2.4.1 Models of Personal Information Management

A number of models of personal information management were discussed earlier in this chapter. **Section 2.2** described a variety of classifications of organising strategies, which are summarised in **Table 1** below. Barreau’s and Boardman’s frameworks of PIM activities were discussed in **Section 2.1.1.2**, and the concepts of reminding vs. finding were discussed in **Section 2.2.4**.

Table 1: Classifications of organising strategies

Reference	Type of information	Classifications
(Malone, 1983)	paper documents	neat, messy
(Mackay, 1988)	Email	prioritizers, archivers, requesters and responders
(Whittaker & Sidner, 1996)	Email	no-filers, frequent-filers, spring-cleaners
(Bälter, 1997)	Email	folderless cleaners, folderless spring-cleaners, cleaners, spring-cleaners
(Gwizdka, 2004)	Email	cleaners, keepers
(Abrams et al., 1998)	web bookmarks	no-filer, creation-time filer, end-of-session filer, sporadic filer
(Boardman & Sasse, 2004)	documents, email and web bookmarks	pro-organising, organising neutral

Another model previously discussed (in **Section 2.2.4**) was the model of information types. Cole (1982) called these ‘action information’, ‘personal work files’ and ‘archive storage’, Barreau and Nardi (1995) called them ‘ephemeral’, ‘working’ and ‘archived’, and Gwizdka (2000) called them ‘ephemeral’, ‘working’ and ‘retrospective’.

These models are all largely descriptive in nature. Few predictive models have been developed, with the only one known to the researcher being Bälter’s (2000) keystroke level analysis of email. This research developed a model for the time it takes to do various tasks such as filing a message and

searching for a message, and then estimated the time taken for these tasks for various different combinations of number of messages and number of folders.

For very small amounts of email (approximately 2 messages per day, 1 search and 50 stored messages), manual search is the most efficient retrieval mechanism, and it is best to have no folders. For larger amounts of email (approximately 10 per day, 4 searches and 1000 stored messages), it is reasonably efficient to use a manual search if around 5-10 folder are used, but it is more efficient to not use any folders and use search tools. For still larger amounts (40 per day, 4 searches and 5000 stored messages) it is much more efficient to use search tools and keep everything in the inbox. Overall, this research concludes that the best long-term strategy for email is to use a few folders in conjunction with the search function. It is also reasonably efficient to use between 5-25 folders and use manual searches.

One limitation of this study is that it greatly depends on the strength of the search function. Also, some emails are difficult to recall by search function because their key topic words are not actually included in the email. This also makes no allowances for hierarchies of folders which might make the browse (and search) tasks easier.

2.4.2 Theories of Workspaces and Distributed Cognition

A workspace is an environment where the cost structure of the information is tailored to the needs of the worker (Robertson, Card, & Mackinlay, 1993). This is true whether talking about a chemistry lab, a carpenter's workshop or a knowledge worker's computer. The workspace needs to be tuned to the tasks the worker is currently performing, so that the information they need is readily available. Thus, there will be sets of information that are frequently accessed and need to be more easily accessible than infrequently used archived information. For example, if the worker is writing a report, the source documents, fact and figures and reference books are all needed close at hand during the task. Once the task or project has ended, these may be filed away. There is a cost in time and effort involved in any action taken in a workspace, and the properties of the workspace determine the structure of costs incurred in taking various actions.

Distributed Cognition is a theoretical framework for modelling cognitive processes that are spread between people, over time and through workspaces (Hollan, Hutchins, & Kirsh, 2000). Some domains to which it has been applied include call centres, air traffic control and ship navigation. Distributed Cognition requires ethnographic field studies with a view to identifying the interdependencies between people and the artifacts in their workspace. A focus is on the correspondence between internal and external representational structures, and an analysis of how cognition is offloaded onto these external representations. Dix identifies two key concepts: (1) triggers, items in the workspace which serve as

reminders that can initiate actions, and (2) placeholders, arrangements of items which can maintain the state of a task so that it can be resumed at a later time (Dix, Wilkinson, & Ramduny, 1998).

External representations, being part of the workspace, have a cost structure associated with their use. Representations used in a workspace are chosen and changed in order to reduce the cost structure of operations in an information processing task (Russell, Stefik, Pirolli, & Card, 1993).

2.4.3 Theories of Classification

Current document management practices rely heavily on classification. As **Section 2.3** showed, most of the prototypes developed in the field of personal information management have been experimenting with different types of classification. This section briefly reviews current theory on classification from the fields of philosophy, linguistics, library science and psychology, and then describes in detail the nature and limitations of the dominant hierarchical classification scheme.

One of the earliest theories of classification began with Plato in ancient Greece (Eysenck & Keane, 1990). This theory holds that each concept has a definition, and matching something to a category is a process of comparing the properties of the object to the defined features of the concept. A more recent theory is Prototype Theory, which maintains that an object is matched to a category according to how similar it is to some exemplar object (Rosch, Mervis, Gray, Johnson, & Braem, 1976). In Prototype theory, category creation is influenced by cognitive economy – people try to maximise the amount of useful knowledge they gain with the least cognitive effort. Lakoff makes the argument based on linguistics that categories are not dependent on objective properties of the object being categorised. Rather they are more subjective, dependent on both the physical and cognitive aspects of people, and are culturally and situationally dependent (Lakoff, 1987).

Bowker and Star (1999) set out three properties of a theoretically ideal classification structure:

1. A consistent, unique classification principle (e.g. temporal, alphabetical)
2. Mutually exclusive categories (everything belongs in only one category)
3. Collectively exhaustive categories (everything can be assigned a category)

No real-world classification matches this structure however, with many schemes embodying multiple principles, multiple categorisations and unclassified items (e.g. ‘miscellaneous’).

Another type of classification structure developed in the field of library science is a faceted classification, developed by Ranganathan (Bowker & Star, 1999). In this type of classification, it is not necessary to choose one unique organising principle. Instead, it is possible to define multiple orthogonal dimensions (or facets) of an item. These facets can then be used to locate an item along multiple dimensions simultaneously. Hierarchically organised facets have been used in some information retrieval systems (Hearst et al., 2002).

Lansdale analysed personal information management and classification from a psychological perspective, and noted that classification of personal information differs from the sort of categorisations performed by a librarian: people do not place each piece of information in a single, unambiguous location, but rather categorise by interpreting the context in which the information appears and is used (Lansdale, 1988). Jones and Ross (2006) suggest that the organising system that a person has in place affects the way new information is perceived and understood, so the type of classification systems that people create are very important, regardless how far they may be from the ideal classification.

2.4.3.1 Classification dimensions

There are a number of candidates for a primary classification principle to be used in personal document management. The most common principle in use is a hierarchical organisation (**Section 2.3.2**). Spatial and temporal dimensions have also been implemented in prototype systems (**Sections 2.3.4 and 2.3.5**). Other systems allow users to avoid having any primary classification at all, but rely on search over user-defined document properties to retrieve items.

There are many other dimensions which could potentially be used as a primary classification principle, including the dimensions noted by Kwasnik (1989) and Gonçalves and Jorge (2004) (**Sections 2.2.1 and 2.2.4**). These include people, place, purpose, use, subject and format. Apart from format, none of these are currently captured by document management systems. According to the subjective classification principle, file format isn't recommended as a primary classification principle (Bergman et al., 2003): *"all information items related to the same subjective topic should be classified under the same category regardless of the technological format"*.

While some of these dimensions may be able to be inferred by a document management system, some are subjective and would have to be supplied by the user.

Spatial location was a very strong cue for organising and retrieval in paper document collections (Cole, 1982; Lansdale, 1988; Malone, 1983). Locations can be strongly associated with particular meanings and humans are able to navigate spaces with relative ease. However, the computer screen is a very limited surface on which to work spatially, especially when people are more used to operating in a 3D environment. Also, spatial organising is subject to the limits of the spatial metaphor, for instance, any piece of information can only occupy one space at a time, and any particular location can only be occupied by one piece of information.

Temporal is one possible organising dimension, however most people do not have a sufficiently well developed sense of time for it to be very useful (Lansdale, 1988 p.61).

Current information workspaces don't privilege one particular categorising dimension above any others. Instead, what they provide is a generic hierarchy of containers which the user can label as they

wish. In this paradigm, each piece of information must be given an explicit name, and must be located in a single, explicitly defined container. The containers (often called folders, directories or nodes) are hierarchically structured.

Setting up a hierarchy of folders is essentially equivalent to defining a set of attributes or keywords that can be applied to a document. For example, consider what it means to place documents into the Lectures folder in the structure shown in **Figure 18** below:

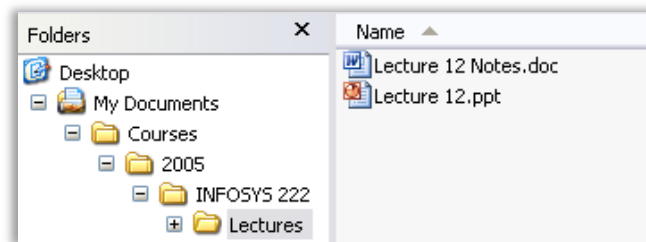


Figure 18: Example Folder Structure

By simply placing the document in this folder, the user is saying that this document is related to their Courses, that it is applicable for 2005, that it pertains to the course INFOSYS 222, and that the document is to do with lectures. These four pieces of information can be assigned to any document with the single action of placing it inside that folder. The folder names provide the context, and the file names and file formats serve to distinguish among documents within that context.

Thus, the folder names that are used, and depth of the folder structure tell us something about the type and quantity of metadata that must be used in order to differentiate documents for use by a particular person.

When people talk about their documents and folders, they often talk about ‘where’ they put things, which can be interpreted as speaking spatially. However, despite the use of language that colloquially indicated spatiality, they are in fact hierarchically instead. They don’t talk about how close or far away our documents are, how high or low, how far to the left or right. Instead the concept is containment—they talk about something being inside or under. It is also tempting to attribute the ability to know that something is in ‘the bottom folder’, or ‘the second one down’ to spatial memory. However, this is just habit; although possibly so well entrenched that the sequence of mouse movements has become part of muscle memory.

No research has yet been done that has looked at how people organise documents and folders into hierarchies.

2.4.3.2 Problems with classification

One of the primary difficulties in the process of filing information is that it is difficult to specify a single, unambiguous classification for any piece of information (Kidd, 1994; Lansdale, 1988; Malone,

1983). Most documents that knowledge workers deal with are complex, encompassing a number of topics, and may have many other characteristics. Trying to determine the single best place in a filing system to locate these documents is a very difficult cognitive task. The difficulty is exacerbated by the fact that the information needs of users change over time, therefore a document that was initially filed in a 'memos' folder may need to be retrieved when searching for all information about the Smith project. It is impossible to predict all possible future uses of a document in order to categorise it appropriately.

Lansdale (1988) identified *"a general problem in categorising items, both in terms of deciding which categorisations to use, and in remembering later exactly what label was assigned to the categorisation"* (p. 55). It is difficult to generate a filing system with simple labels that will cope with all types of information. There are two reasons for this: (1) it is difficult to come up with a comprehensive set of unambiguous category names to use, and (2) information usually does not neatly fit into a single well-defined category, often falling into several categories simultaneously. The constraint of forcing information into a single category means that the system is unable to cope with changes over time. Any given piece of information may need to be recalled in a different manner in the future, and single categorisation does not support this (Lansdale, 1988).

Malone (1983) identified both cognitive and physical barriers to effective classification, noting that mechanical difficulty in establishing a labelled filing system should be alleviated by computer-based systems making creation and manipulation of categories easier, leaving only the cognitive difficulties to deal with. This cognitive difficulty is so great that many users do not even attempt to classify documents at all, fearing that if they file them, they will be 'lost' (Lansdale, 1988; Malone, 1983). Some of the suggested ways of overcoming this barrier include automatic classification, allowing multiple classifications, and having no initial naming or classification required of documents. The latter suggestion mimics the ease of use of paper, since people can begin writing on a sheet of paper without giving it a name or specifying what folder it will be filed into.

While items often remain unclassified due to the cognitive and practical barriers of classification, there are additional reasons why an item may not be placed into a classification. These are (1) so that the document can serve a reminding function, and (2) in order to keep active documents close at hand during a task (Lansdale, 1988).

2.4.3.3 Automatic classification

While automatic classification has been suggested as a partial solution to cognitive difficulties with filing, it is not without its own problems. Less cognitive involvement in filing means that a user is less likely to remember how an item was filed, and therefore both less likely to be able to recall it, and less likely to recognise it (Lansdale, 1988 p.65). This partly explains why the cognitive load incurred when

filing documents is necessary, since without it, the user may not even remember the existence of the document and therefore won't even initiate a search for it.

Intended or current use of a document proved to be one of the primary classifiers of a document, and Kwasnik noted that "*classification is, overall and above all, person- and situation-centred, and not object centred*" (Kwasnik, 1989, p. 207). This would seem to indicate that automatic classification may not be possible. Research that has examined automatic classification in email (Ducheneaut & Bellotti, 2001; Whittaker & Sidner, 1996) found that people tend not to use rules, but process the messages manually and retain control over where they are filed.

It is generally known that memory is better for self-generated material than for externally imposed categories (Lansdale, 1988). Thus it would be easier to locate material for which the filing category had been decided by the user, than if the categorisation was done automatically. MailCat (Segal & Kephart, 1999, 2000) offers a promising compromise, presenting the user with several classification possibilities, and allowing one to be chosen with a single mouse click.

2.4.4 Theories of Information Retrieval

As noted in **Section 2.1.1.4**, Information Retrieval is a significant part of personal document management. This section reviews theories of information retrieval that are applicable to personal document management.

Lansdale proposes a framework for information retrieval: "*every attempt at retrieving information involves two distinct psychological processes: recall-directed search, followed by recognition-based scanning*" (Lansdale, 1988, p. 64). The process of retrieving information from an information system is a trade-off between these two activities. It is known that human recognition ability is far superior to recall. In fact the human ability to exactly recall information (particularly arbitrary information) is quite limited (Miller, 1956). Thus, information workspaces need to support recognition-based processes as well as recall-based processes (Baecker, Booth, Jovicic, McGrenere, & Moore, 2000). People tend to remember the general subject and meaning of items and events, but do not have a good memory for the details. They often are able to remember more details than they are able to recall:

"As it stands, therefore, there is good reason to believe the informal observations that what is remembered about documents is the meaning of their content and contextual information such as what they looked like, what one was doing at the time, and so on. As we shall see, computer systems do not use this information but rely upon the user remembering filenames and/or the categorisation of required information. What we are good at is being ignored, and what we are required to do is a difficult and flawed psychological process". (Lansdale, 1988, p. 59)

Another trade-off involving retrieval is between the effort expended in filing the information initially, and the effort required to retrieve it later. It is assumed that the main purpose of creating and

structuring a filing system is so that the contents may be easily retrieved when required. The total cost (in terms of time and effort) required to use an information workspace is a combination of the time spent filing (storing) the information, and the time spent finding (retrieving) information. If more time is spent in establishing a useful classification system, and time is spent carefully filing each document into the system, it is likely that retrieval will be fast. However, if almost no effort is spent on filing, the searching for documents is likely to be rather slower.

"The more we ask the user to do at the process of information storage, the less likely he is to do it, creating retrieval problems. On the other hand, the more we automate the process of storage and take responsibility away from the user, the less he is going to remember, and therefore the less he is going to be able to retrieve." (Lansdale, 1988, p. 65)

Spending a great deal of time setting up an elaborate filing system that speeds retrieval may not be useful for someone who only infrequently retrieves documents. However, someone who frequently needs to access archived material will have a lot of trouble if everything is stored in one huge box. The document management strategies described in **Section 2.4.1** can be considered to be taking a position with respect to this trade-off. **Figure 19** below shows the relative amounts of effort spent in filing and finding for 'pilers' and 'filers':

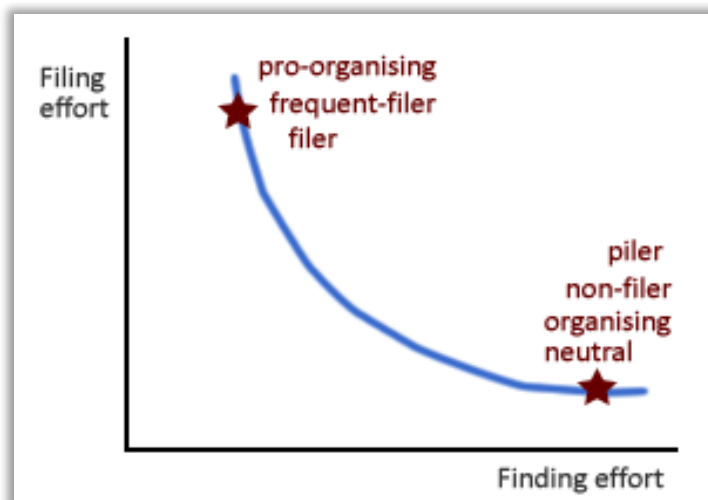


Figure 19: Trade-off between filing and finding

One of the assumptions made in much of the earlier research is that people who spend more time filing up front are able to locate their documents more easily and hence the up-front investment in organising pays off. However, it is not clear whether or not this is true. Predicting the optimum amount of organising to do requires knowledge of future retrieval requirements. This can never be perfect, but it is reasonable that over time, people would get better at knowing what is required. Thus, a person's strategies and structures may converge on an optimum solution given their patterns of document use. In addition, some personality factors have been postulated to affect where on that continuum people feel most comfortable. In particular, Gwizdka identified flexibility of closure as one

important psychological property that affects which strategy people are likely to adopt (Gwizdka, 2004). To date, no longitudinal studies have been done comparing how much time people adopting various strategies spend in information management.

It is assumed that the major motivation for an individual to establish an elaborate filing system is to aid retrieval, and that the process of filing is in effect a coping strategy that allows the individual to find information in the absence of any other means of doing so. Given the increasing sophistication and ease of use of information retrieval technologies, it is feasible to wonder whether structuring and filing of information might eventually be rendered obsolete by a sophisticated search facility. This would be the case if the structuring and filing of information served no other purpose than facilitating finding. It is possible that information management may be more complex however, and that the act of classifying and organising information might be valuable in itself. However, it is likely that changes in information retrieval technologies will alter the position of the trade-off between filing and finding. **Figure 20** below shows one possible consequence of improved indexing and search user interfaces. Notice that the shifting curve benefits non-filers significantly more than filers.

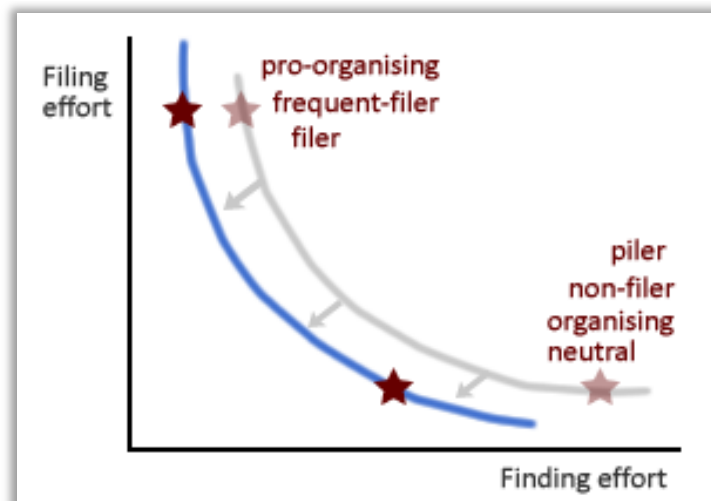


Figure 20: Trade-off between filing and finding shifting with improved search

This will make it possible for people to profitably adopt a different position on the curve, an example of the Task-Artifact cycle in action. However, at this stage, not enough is known to predict how this curve will actually change.

2.5 CONCLUSION

This chapter has reviewed the literature relating to personal document management. **Section 2.1** defined the term personal document management as the activity of managing a collection of digital documents performed by the owner of the documents. Management includes creation and acquisition, retrieval, organisation and maintenance. The study of personal document management was placed within the research domain Human-Computer Interaction and located with respect to related fields of

inquiry. It was noted that personal document management is an embedded and intermittent activity. **Section 2.2** reviewed the empirical studies that have been conducted on personal information management, including studies on paper documents, email and web bookmarks as well as digital documents. It was noted that there is a distinct lack of recent empirical research into personal document management, and that further work is warranted in this area.

Section 2.3 surveyed some of the most influential commercial systems and research prototypes that have been developed to support document management, noting very little systematic research has been conducted into how these are used and what problems users might encounter with them. Most are not based on empirical research into personal document management.

Section 2.4 reviewed the theory relating to personal document management. Existing models of personal information management were reviewed, and again the lack of research concerning document management was noted. The concepts of workspaces and distributed cognition were presented as possible models for explaining the use of document collections. This was followed by a review of theory related to classification, including a discussion of the cognitive difficulties users have with classification. Theory relating to information retrieval was also reviewed, with the many classifications of personal information management behaviour being conceptualised as positions adopted with respect to a trade-off between filing and finding.

This review of current document management literature points to a clear need for further research into personal document management. In particular, further research is needed in these areas:

- How individuals manage their personal documents
- How individuals structure their personal document collections
- What problems individuals encounter with document management activities
- How well existing tools support document management
- Requirements for effective tool support of document management

Without empirical information about personal document management, it is difficult to effectively develop supporting tools and to evaluate those tools. This research aims to fill this gap in the current research landscape.

Chapter 3

RESEARCH DESIGN

The previous chapter concluded with the current state of knowledge in personal document management, and identified a number of areas where more research is necessary. This chapter begins with a summary of the specific research questions that need to be answered in order to advance knowledge in this area.

Next, the design of the research is described a number of levels. The first step is to determine the overall type of research that is to be carried out, and then select a research strategy (or strategies) appropriate to the phenomenon under study and to the research type. Next, within each research strategy, the data collection techniques must be chosen, and finally, the participants for the study will be identified.

Because multiple research strategies have been selected, the specific details of the design, implementation and analysis of each strategy are deferred to chapters dealing with each strategy separately.

3.1 RESEARCH QUESTIONS

From the preceding chapter, it is clear that the most significant gap in the research to date is lack of any recent information (and particularly quantitative information) about personal document management strategies and structures. These are the specific research questions that this research aims to answer:

What strategies do people adopt in order to manage their personal document management?

What kinds of structures do people create to organise their documents?

What problems do people have with the current tools for supporting document management?

Due to the lack of prior research into these areas, the nature of the research is necessarily fairly exploratory.

3.2 TYPE OF RESEARCH

Research can be conducted for several distinct reasons or purposes. These are often labelled exploratory, descriptive and explanatory (Pinsonneault & Kraemer, 1993; Yin, 1989). Other writers categorise the research types into exploratory and confirmatory (Boudreau, Gefen, & Straub, 2001; Hair, Anderson, Tatham, & Black, 1995) or into descriptive and explanatory (de Vaus, 2001). These categorisations reflect different ways of partitioning a continuum of types of research. The classification of exploratory, descriptive and explanatory will be used for the purposes of placing this research.

Exploratory research is done when there is little or no research existing in a given field, and the purpose of it is to better understand the nature of the problem (Sekaran, 2000). It aims to uncover potentially important characteristics of a situation and to identify areas where further study may be necessary. Descriptive research aims to describe the range of situations, opinions or events that can occur and how those are distributed amongst the population (Pinsonneault & Kraemer, 1993). It is often used to identify and define possible relationships at a fairly general level. Explanatory research is aimed at exploring relationships between things, testing theories about those relationships and possibly inferring cause-and-effect. It is concerned with answering the question of why things happen the way they do (de Vaus, 2001).

Given the state of current knowledge in personal document management and the research questions to be answered, this research can be classified as partly exploratory and partly descriptive. Some aspects of document management have been studied so little that exploration is needed to even identify issues that are relevant. However, there are some concepts raised by prior research and by research on personal information management topics that require more detailed information to be gathered.

3.3 RESEARCH STRATEGY

Research Strategy refers to the overall architecture of the research project. It is concerned with deciding the high level features of the study, rather than the details of the techniques used and implemented. It 'deals with a logical problem and not a logistical problem' (Yin, 1989).

Many researchers have provided frameworks enumerating the different research strategies available (e.g. Alavi & Carlson, 1992; Boudreau et al., 2001; de Vaus, 2001; Jenkins, 1985; Orlikowski & Baroudi, 1991; Yin, 2003), and some have provided guidelines about the appropriate circumstances in which to

choose each strategy (e.g. Yin, 2003). Since the aim of this research is to gather empirical evidence concerning personal document management practices, only empirical approaches will be considered here.

Table 2 shows the major research strategies identified by four authors, and maps them together into a common framework. The definitions of these strategies that follow are taken from Boudreau et al (2001) except where noted.

Table 2: Comparison of Research Strategy Frameworks

Research Strategy	Yin (2003)	Alavi and Carlson (1992)	Boudreau et al (2001)	Orlowski and Baroudi (1991)
Experiment	Experiment	Lab experiment Field experiment	Laboratory experiment Field Experiment	Laboratory experiment
Survey	Survey	Survey		Survey
Field Study		Field Study	Field Study	
Case Study	Case Study	Case Study	Case Study	Case Study

Experimental research is conducted in a setting largely controlled by the researcher. Participants are usually randomly allocated to two or more groups which differ in some aspect of the phenomenon being studied. All other elements that may vary are controlled as well as possible by the researcher. Laboratory experiments occur in a setting created by the research for the purpose of the research, whereas field experiments are situated in the natural environment of the phenomenon being studied. Experiments are generally used for confirmatory research, with results generally being derived by finding differences between the groups under study. The use of statistical techniques means that experiments usually require a medium to large number of participants.

Surveys are usually situated in natural environments where the researcher cannot usually control or manipulate any aspect of the situation under study. Alavi and Carlson (1992) and Pinsonneault and Kraemer (1993) consider survey to be a strategy in its own right, but this research follows Boudreau et al. (2001) in considering surveys to be a particular data collection technique, rather than a research strategy. The term questionnaire is used to denote the data collection technique that is usually associated with surveys. However, surveys don't always use questionnaires, and questionnaires may be used with other research strategies. Surveys are characterised by collecting information from a relatively large number of subjects using sampling logic and statistical analysis to generalise the results to a population. Surveys are not well suited for exploratory research on their own, since they typically collect information using a fairly structured set of questions (Pinsonneault & Kraemer, 1993). Thus they

are better suited to descriptive or explanatory research. Because of the use of statistical techniques, a medium to large number of participants is usually required.

Case studies intensely examine a very small number of situations in great detail. There is no ability to manipulate or control for aspects of the setting. They may use any kind of data collection technique, and usually use multiple techniques and typically generate a large amount of data. They rely on analytical generalisation (generalisation to theory) rather than statistical generalisation (generalisation to a population). Because of their resource intensive nature, the number of situations studied is usually very small, often only a single situation.

Field studies occupy a middle ground between case studies and surveys. They are situated in a natural environment where the researcher cannot control or manipulate anything, and there is no control group. Field studies may employ many different data gathering techniques, including interviews, observation and questionnaires, and often combines more than one. Boudreau et al note that ‘sometimes researchers will refer to “multiple” case studies, which, when they exceed a dozen or more sites, are more than likely classifiable as field studies. Field studies typically have fewer participants than surveys, and usually do not rely on statistical significance for their results.

For this research, the primary considerations are the desire to be able to gather data about personal document management practices from a larger number of participants, and the necessity of capturing information about personal document management embedded in the real world context. The need for a large number of participants rules out case studies, and the need to preserve the natural environment eliminates experiments.

A survey strategy would provide the necessary ability to generalise the results to a wider population, and the strategy is particularly suited for a large number of participants. Although a survey technique can be used for exploratory research, it is best suited for descriptive or explanatory studies due to the need to have a structured set of questions. The lack of knowledge in the field of personal document management at the moment means that it isn’t possible to generate a set of questions that would adequately cover the range of issues in personal document management, since the range of issues is not fully known.

A field study would allow the investigation of personal document management to be situated in its natural context, and do not typically require that all issues to be investigated are fully known in advance. However, field studies are typically more time-intensive and generate a lot more data than surveys, and especially so if done with a number of participants that is sufficiently large to give generalisable results.

Since neither of these strategies on their own is ideal, they will be combined in order to attenuate each other’s weaknesses. The research will start with an exploratory field study with a small number of

participants, designed to explore the issues of personal document management in the current workplace. This will generate qualitative data that can be examined to generate a conceptual framework of the issues. This conceptual framework can then be used to generate a set of structured questions that can be used in a descriptive survey to get quantitative data that can be generalised to wider populations. **Figure 21** shows the overall architecture of this research.

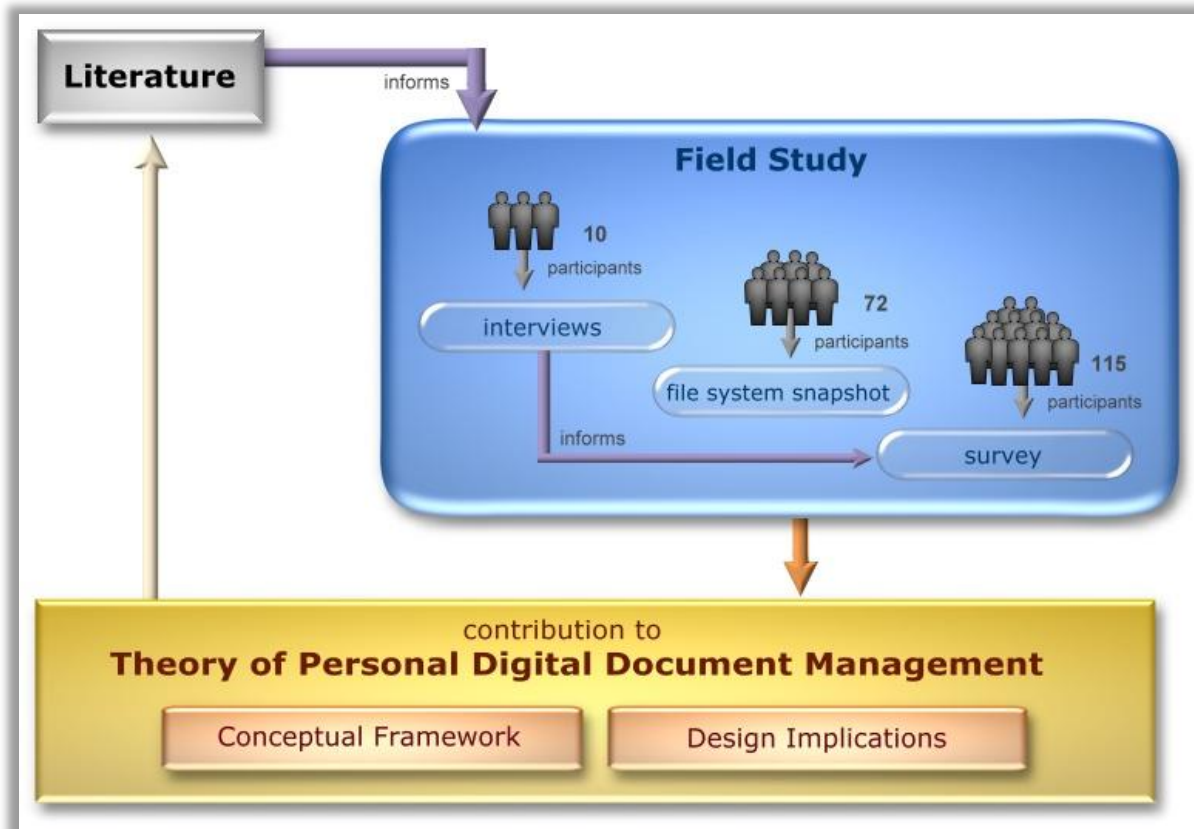


Figure 21: Overall field study research strategy combining interviews and survey

Combining multiple strategies in this way is often referred to as a multi-methodological approach. This approach is recommended in information systems research, since no single research strategy is believed to be sufficient alone (Nunamaker, Chen, & Purdin, 1990). Taking a multi-methodological approach increases the number of perspectives from which the phenomenon can be studied, and increases the possibilities for gaining understanding (Pinsonneault & Kraemer, 1993).

3.4 DATA COLLECTION METHODS

There are a large number of data collection techniques that can be used in research: questionnaires, interviews, archival documents, direct observation, participant observation, measurements, examination of artefacts and others.

A combination of interviews, questionnaires and examination of document management artefacts was selected for the two phases of this research. The following two sections explain the data collection methods employed in the field study and the survey.

3.4.1 Interviews and File System Snapshot

It was decided to use face-to-face interviews as the primary data collection technique for the first part of the field study. Interviews allow the collection of subjective information from human subjects. Interviews are useful for gathering information about opinions, internal decisions and thought processes as well as other features of human cognition that cannot be observed directly. The interviews can capture people's self-reported experiences of document management, which reflects the person's own perceptions of what they do. Face-to-face interaction also allows the interviewer to observe the physical environment in which people are working, and to refer to artefacts of the document management activity during the interview.

This technique of interviewing participants in their offices and using their computers as a questioning point for the interview has been used many times in investigation of related aspects of personal information management (Ducheneaut & Bellotti, 2001; Malone, 1983; Whittaker & Hirschberg, 2001; Whittaker & Sidner, 1996), and was used in the only prior studies of personal document management (Barreau, 1995; Barreau & Nardi, 1995).

One limitation of interviews is that self-reported experiences are often not accurate. There are a number of reasons for this, including imperfect memory (Schacter, 1999), wanting to give an answer that is socially acceptable (Schwarz & Oyserman, 2001) and making different assumptions to the researcher about what is important (Schwarz, 1999). These issues have also been found in usability studies, where what people say they do does not always exactly match what occurs (Nielsen, 2001).

For these reasons, it is useful to combine the subjective data collected in the interviews with some objective data about actual structures and practices. There are two possible aspects that could be investigated here – the document management activities, processes and behaviour, and the document structures that are created. Objectively investigating the activities performed would require a fairly extensive period of observation, either directly, through video recording, or through software instrumentation, in order to find out what activities actually occurred. While this is useful research, and should be performed, it is beyond the scope of this research.

Therefore, it was decided to collect data about the actual document structures that were used, and combine that with the interview data about the structures and the activities and behaviour that surround them. To do that, a snapshot will be taken of the file structure of each field study participant.

This snapshot will provide data about the number and types of folder and files and how they are structured together.

This quantitative data about the structure will be combined with the qualitative data from the interviews about how the structure is used. A number of researchers have argued that the combination of both qualitative and quantitative methods results in more robust research, and can allow more complex and novel explanations (Kaplan & Duchon, 1988; Mingers, 2001; Petter & Gallivan, 2004).

The design of the interviews and snapshot are described in **Section 4.1**. This includes details of the design and implementation of both the file system snapshot and the interview protocol. The data analysis strategy for those data collection methods are also described there, as are the steps taken to ensure the reliability and the validity of the results.

3.4.2 Survey

The purpose of the survey in this research is to gather descriptive data in order to be able to better describe the personal document management landscape. The survey shares with the field studies the need to collect both subjective data about document management practices, as well as objective data about document structures. For this reason, the survey will also include a file system snapshot. However, because the aim of the survey is less exploratory and more descriptive, the survey will use a structured questionnaire. This will allow quantitative analysis of the survey results, and will also enable data to be collected from a larger number of people.

Because the survey is going to be informed by the results of the field studies, the development of the survey questionnaire will be discussed in **Section 5.1.1**.

3.5 PARTICIPANT SELECTION

The context of personal document management is a person and their computer, and therefore this is the natural unit of analysis. While people engage in document management practices in a home setting, and for leisure purposes, it was decided to focus on the work setting because the work situation is likely to be where this subject is most germane. For instance, it is in a work context that the issue of information overload has been raised, and it is in a work context that the issue of productivity declines as a result of ineffective information retrieval have been identified (Farhoomand & Drury, 2002). It is expected that the work situation is the highest use, and has the most negative consequences if document management practices are inefficient.

These days it is becoming increasingly more likely that people do not just use a single computer exclusively, but may use a range of devices. However, it is still a widespread practice for most employees in an organisation to have a primary computer that is 'theirs', on which they store their

information. This leads to a narrower unit of study consisting of a knowledge worker and their primary computer.

Within the population of knowledge workers in a work setting, it was decided to restrict the study to users of the Microsoft Windows operating system. This was partly for practical reasons, since different operating structures store files in different ways, and obtaining tools to take a snapshot from any operating system is likely to be extremely difficult. The decision was justified by the fact that the Windows operating system has approximately 93% of the desktop operating system market (MacInnis, 2003), and is more prevalent in organisations than in home settings. In addition, the other major software systems, Apple Macintosh and GNU/Linux have hierarchical file browsers very similar to Windows Explorer (and in some cases almost indistinguishable from it). However, it is possible that users of these other operating systems (particularly users of Apple Macintosh) may have different work practices to Windows users, and therefore the recommended user interface guidelines produced by this research may not be applicable to them.

Within the populations of knowledge workers using the Windows operating system in a work context, the selected sampling frame is the staff members of the University of Auckland Business School. This was partly a matter of convenience, since the researcher has easy access to them and they have a standardised hardware platform running Microsoft Windows. One issue raised by prior research is that there may be a difference in the document management needs of researchers and non-researcher knowledge workers. The Business School employs staff in both categories, with approximately 442 staff members listed in the staff directory total. Of these 237 were academic staff, and 205 were classed as general staff. Both the academic and administrative staff fit the definition of knowledge worker given by Drucker (1959) and therefore all were included in the study.

The Business School staff includes people from a wide variety of backgrounds, many different countries, and a wide range of ages and experiences. The academic staff roles typically involve both research and teaching duties, although some staff have only teaching duties, or teaching administration duties. The general staff include managers, personal assistants, departmental administrators, secretaries, and technical services.

3.5.1 Generalisability

Although this study is conducted with participants using a single operating system, because most operating systems provide the same basic functions for document management, it is likely the findings would be applicable to other operating systems such as Linux and Apple Macintosh. The Business School is fairly permissive in giving employees full control of their work computers, including the ability to install any software of their choice. Thus, it is expected that the population will be reasonably

representative of the wider population in their ability and inclination to try tools such as Google Desktop or Copernic Desktop Search. As people may differ in their approach to document management in other settings, the findings may not be able to be generalised beyond the work context, and since the research is conducted with people using PCs, it is not likely to generalise well to other devices or interaction modes. The diversity of backgrounds and mix of ages and genders found amongst the Business School staff should ensure the findings are applicable to most demographics, and the inclusion of general staff as well as academics ensures the results can be generalised beyond an academic setting.

3.6 ETHICAL CONSIDERATIONS

The use of human participants in research requires the approval of an ethics committee, to ensure the research complies with ethical guidelines. The requirements of the ethics committee were that the participants should have the right to refuse to participate in the research without explaining why and that they should have the ability to withdraw from the research at any time. The ethics guidelines also dictate that participants should be fully informed of the research that they are undertaking and they need to be told exactly what is required of them. Further, the ethics guidelines suggest that questionnaires should be anonymous unless there is a compelling reason for them not to be.

3.7 CONCLUSION

Two research strategies have been selected to answer the research questions surrounding personal document management. The first is an exploratory field study that will generate both qualitative and quantitative data about personal document management structures and practices, through the use of face-to-face interviews and a file system snapshot. The interview data will be analysed to generate a conceptual framework of personal document management issues and behaviours, and this framework will be used to inform the second research strategy: a survey. The survey will be a cross-sectional, descriptive survey, designed to provide a snapshot of the distribution of various behaviours, structures, problems and opinions on personal document management in current workplaces. The survey will use both a structured questionnaire and a file system snapshot in order to combine subjective information about document management behaviour with objective information about document structures.

The next two chapters explain in more detail how each of the three data collection techniques were developed and administered and present the results that were obtained with each instrument.

INTERVIEWS AND FILE SYSTEM SNAPSHOT

The previous chapter outlined the overall design of this research and identified the role of the field study in gathering exploratory information about personal document management practices and structures. This chapter describes the initial phase of the field study, encompassing the design, implementation and results of both the interviews and the file system snapshot results.

Section 4.1 describes the interview protocol used and the file system snapshot software that was used to gather empirical data. **Section 4.2** presents the results obtained by the interviews. This includes summaries of the document management behaviour of each participant, followed by the exposition of an initial conceptual model encompassing the themes raised during the interviews.

The results from the file system snapshot are presented in **Section 4.3**, including a discussion of the metrics that have been developed to describe file system structures. This section also includes an analysis of folder naming practices. Finally, **Section 4.4** integrates the qualitative results from the interviews with the quantitative results from the file system snapshot in order to propose three distinct document management strategies.

4.1 STUDY DESIGN

This section describes the design of this study. **Section 4.1.1** describes the selection of research methods adopted in this study, and provides a justification for this choice. **Section 4.1.2** describes how the study participants were selected, and **Section 4.1.3** describes the interviews, including the evolution of the interviews through the study.

4.1.1 Choice of Method

Interviews can range from fairly structured affairs where a questionnaire is administered face to face to a very unstructured and rich interaction which verges on being a mini ethnographic study. The more ethnographic style interview has the advantage of being able to access more context and to be guided by the particular circumstances of the participant, but because of the very depth and richness, it generates a lot more information to analyse and thus the number that can be performed is more limited. Many of the studies discussed in the review of theory have used a form of ethnography (sometimes called 'ethnography lite') in order to investigate this situation. The essence of this technique is conducting in-depth unstructured or semi-structured interviews in the context under investigation. This allows the contextual nature to be considered, and allows issues and concepts to emerge naturally. This technique has been successfully used in a number of investigations of different aspects of personal information management, and is well suited to this topic.

Because of the contextual and embedded nature of document management, it was decided to begin with semi-structured ethnographic interviews. These interviews are carried out in the workspace of the participant, in which the participants are encouraged to demonstrate their work practices to the researcher rather than simply explaining them. The semi-structured nature of the interviews means that (although guided by an initial framework) the concerns and processes of the participants are able to emerge.

Although contextual information from the environment is included, the main source of data is the self-reported perceptions and opinions of the person involved. This is very good for understanding how they view the issue and what problems they encounter, but cannot provide information about exactly what they do, and what structures they create. Objective information about their document structures will be collected through use of a file system snapshot tool.

These interviews will then be used to develop a conceptual framework of the issues surrounding personal document management, and this in turn will be used to guide the development of a survey instrument which will collect information from a wider range and number of participants.

4.1.2 Participants

The participants were drawn from the staff of the University of Auckland Business School (both general and academic staff).

Because of the perceived privacy implications of examining someone's files and folders, the first four participants selected were personally known to the researcher. The remaining six were selected in order to provide coverage of different departments and different job titles within the Business School. Invitations were sent until the necessary 10 participants had been found. 11 people were invited and 5

declined to take part. As per the ethical guidelines, they were not required to give a reason for declining, but 2 indicated that they felt they were too busy to participate.

The following table summarises the demographics of the 10 participants.

Table 3: Field Study Participant Summary

Code	Role	Gender	Age Range	Interview Duration	Alias
A	General staff	Male	20-29	107 min	Alex
B	Senior Tutor	Male	20-29	37 min	Brett
C	Senior Tutor	Female	40-49	38 min	Candice
D	Senior Tutor	Male	20-29	63 min	Damien
E	Lecturer	Male	50-59	37 min	Edward
F	Lecturer	Male	50-59	31 min	Frank
G	Lecturer	Female	Unspecified	42 min	Gail
H	Senior Lecturer	Female	40-49	28 min	Harriet
I	Professor	Female	40-49	32 min	Ina
J	Associate Professor	Male	60-69	46 min	Jack

The 10 participants came from 4 different departments within the Business School. The participants' departments are not identified as this may make it possible to identify the participants.

Participants have been assigned an alias in order to be able to naturally discuss their interviews without compromising confidentiality.

4.1.3 Interview Process

Participants were interviewed in their own offices so that their document management practices could be viewed in their natural context. The interviews typically took between 45 and 60 minutes. Prior to the interview, the participants were told not to do anything special in preparation for the interview, and particularly, not to clean up their folders beforehand.

At the start of the interview, the research project was explained to each participant and they were reassured that their document management practices were not being evaluated in any way. They were also promised anonymity in the write up. Two participants asked that no direct quotes or direct depictions of their file system be included in the results, even if anonymised. All participants signed a consent form acknowledging their agreement to participate in the research (a sample of this is provided in **Appendix A**).

The interview started by running the file system snapshot software (see **Section 4.2.2.8**), and recording basic demographic information while the snapshot was running. For this purpose, an

Interview Record sheet was used. A sample of this is provided in **Appendix B**. Once the snapshot was completed, the participants were asked to ‘show me where you keep your documents’. The tour of the file system usually centred on the Desktop and My Documents folders, although participants were asked about other locations as well. System created files and folders were ignored, as were files and folders that contained program code and applications.

The interview was largely guided by the participants’ concern and observations. The interview protocol was used as a guide to ensure all topics of interest were covered during the interview. The interview protocol is provided in **Appendix C**. Not all areas of the interview protocol were relevant to all participants. For example, some participants did not use the Desktop in any way and so they were not prompted further about the ways they use the Desktop.

Participants were often asked to demonstrate as well as explain their practices, e.g. ‘can you show me what you would do?’ Participants were also asked about what they liked about the document management facilities in Windows, as well as asked about what they didn’t like.

All interviews were audio-taped and were transcribed as soon as possible after the interview. During the interview, the interviewer took rough notes about the participants’ actions during the interview, and also tried to narrate their actions so there was a record on the transcript. The transcript was augmented with descriptions of the actions from these sources and from memory. For example, the transcript might record that the participant said ‘these files over here are to do with my courses’. From the notes, narration and memory the transcript would be annotated to indicate that the participant was pointing to the lower left of the Desktop. An extract from one of the transcripts is provided in **Appendix D**

During the transcription, the major observations, themes and questions that arose were recorded on a Contact Note. An example of a Contact Note is provided in **Appendix E**. This was used to revise the interview protocol after each interview in order to explore newly observed concepts. The following section explains the evolution in the interview protocol.

After the transcription, a narrative Participant Summary was created which condenses the participant’s background and demographics as well as the salient points of the document management behaviour. The participants were offered a copy of their transcript and this summary to review, although most participants did not wish to do this. The participant summaries are in **Section 4.2.1**.

In summary, for each participant, the following items were collected:

- Audio transcript of the interview
- Snapshot of the file system
- Interview Record Sheet with demographics and system information
- Simplified Transcript of the interview

- Participant Summary of personal document management behaviour and issues
- Contact Note made after transcribing with the major themes raised by the interview and the most significant questions raised for further investigation.

4.1.3.1 Interview protocol evolution

The interview protocol used in this research was in the form of a list of topics to explore during the interview. It did not contain specific questions to ask, nor prescribe a particular order to adhere to. The interview was largely guided by the participant's explanations of their processes.

The initial list of topics to cover in the interview was derived from the literature review in the previous chapter. As the interviews progressed, some of these topics were dropped from the protocol as they were not issues that the participants were concerned with. At the same time, new topics were added as completely unsuspected topics of concern emerged. The interview protocol was adjusted after each interview to further explore the topics that were being raised.

After all topics had been sufficiently covered (either spontaneously or in response to a question), participants were asked about the things that annoyed them the most, wasted their time, and the things that helped them the most in their file system.

In the first two interviews, participants were asked about their ephemeral, working or archived information, and asked for an estimate of the proportion of their information that was either ephemeral, working or archived. This proved extremely difficult for both A and B. This is due to there being no clear-cut distinction between the information types, and also the temporal nature of the phenomena. For instance, much of what is in the email system once was ephemeral, but now is of no use, and will never be referred to again. Is this still counted as ephemeral, or as archived? The difficulty of these distinctions, and the likely variance in assessment by different participants, coupled with the likely limited usefulness of this information led to this question being dropped.

Participants were asked about the good and bad features of the file system at the end of the interview. However, after the first two interviews, it seemed that the issues that were raised through prompting from the interview protocol would be likely to bias the response, and so these questions were shifted to the start of the interview.

4.2 INTERVIEW RESULTS

This section presents the results of the interviews. **Section 4.2.1** provides a summary of each participant's document management practices, and **Section 4.2.2** presents the results of the qualitative analysis of the interviews.

4.2.1 Interview Summaries

The following sections summarise the interviews with each participant. Note that in the descriptions below, participants are referred to using their aliases.

Summary of Interview with Participant A (Alex)

Alex develops and supports software for the Business School. He is in his 20s, and has been in the job full time for one year, and part time for two years prior to that. He believes that his documents and email are very disorganised and that he is not very good at managing his personal information.

Alex's document management activity is driven almost entirely by email. All the documents on his system were either received through email for him to review or update, or were created with the intention of immediately emailing them to other members of his team.

Most files and documents that Alex needs to modify are saved to the Desktop. Items appear on the desktop in columns in the order they were saved there, so the last one on the list is the most recent. He can see how old a document is (relatively speaking) by its position in the list. Periodically, his Desktop overflows. When this happens, he creates a folder on the desktop named cleanup, and drags almost the entire Desktop contents into it. Over time he has accumulated a succession of these folders (named cleanup1, cleanup2 etc up to cleanup10). During a cleanup, these folders remain on the desktop, along with a couple of commonly used shortcuts. He deletes very infrequently, deleting only documents of no intrinsic value to him at all, and when he does so, he permanently deletes them, bypassing the Recycle Bin.

Some documents that Alex perceives are of more permanent value to him and his team get dragged into My Documents when they are complete, rather than remaining on the Desktop, but this is infrequent. Even more rarely, Alex will create subfolders in My Documents to gather together documents related to a specific project. Efforts like this are usually quickly abandoned and the other files to do with the project could end up on the Desktop or in My Documents itself.

Alex makes very little use of subdirectories. In addition, he doesn't use the tree view in Windows Explorer to navigate through his folder hierarchy. Instead, he starts typing the path name into the Address bar and allows the auto complete function to finish the path. He doesn't navigate through the hierarchy using the mouse. He always uses the details view, sorted by either name or date.

Alex creates documents by opening the application, typing the title of the document then saving, and allowing the application to use the title as the default file name. His file names tend to be long and descriptive, and always match the internal document title. Alex doesn't use naming schemes at all, but if a document is specifically tied to a particular date, he will put the date in the title (in year, month, day order to facilitate sorting)

If a document that Alex needs to open is not on the visible Desktop, he will probably use the search function to find it, rather than browse through all the old cleanup folders. He is probably more likely to search for the document in the email system than in the file system, since he knows that the most recent version of a document will always be in his email system (unless he's actively in the process of changing the document, in which case it will be on the visible Desktop). He will usually remember who the document was received from or sent to, and the general topic (such as project name). He may also remember time information, but only very roughly. He may be able to give an approximate time span (e.g., yesterday, last week, a few months ago), but is more likely to recall that a document was worked on after something, and before something else. When he does search the file system, he often searches for key words in the title, knowing that his descriptive file names often provide reasonable accuracy. He will sometimes filter by file type if he knows he is looking for a particular type of file. He will never specify date or size as search criteria, but will frequently sort the list of files by these things (particularly date) to make it easier to locate the item he is looking for.

Alex's biggest problem with his documents is that he has many files that have lost their context. Whereas an email attachment has a sender and some indication of what he needs to do with it, that context is lost in the file system. So Alex has no idea of what he has to do with any given file, and no way of tracking whether or not he has done it. Instead, he uses the email system for this task management. In fact, he often emails files to himself so he can have that context.

Alex makes very little use of folders in his email system, keeping thousands of messages in his inbox. He relies totally on messages being marked as read or unread to know what he still has to deal with. He makes heavy use of the search function when he needs to find messages again, using the Advanced Find interface and usually searching by sender and receiver and by keywords from the message itself, and sometimes by whether or not there is an attachment.

His biggest problem in email is using it to manage workflow and tasks. He has difficulty with the number of ephemeral messages, particular since there is no way to group them into conversation threads.

Summary of Interview with Participant B (Brett)

Brett is a Course Manager at Auckland University. He is responsible for administering a number of very large courses, and his duties include hiring, training and managing staff. In addition, he is currently working on his PhD. He is in his 20s, and has been in the job full time for two year, and part time for three years prior to that. He believes that his documents and email are quite well organised.

Brett keeps his active documents on the Desktop. When the desktop gets cluttered, the files are placed into folders on the desktop. Some of the folders on the Desktop relate to specific courses or activities, but others group documents by file type, e.g. "Word Docs", "Excel Docs", "Access Docs". Most

of the time, the files in these file type folders are on similar topics. Brett also keeps some documents in My Documents. Usually this happens when My Documents is the default place for the document to be saved and Brett doesn't change it. When he does deliberately choose a location for a file, it is usually the desktop.

Periodically (no more than every semester or year) everything on the Desktop or My Documents is moved into a dated folder in C drive. He very rarely has to retrieve anything from these stores, perhaps once or twice a year at most. He could probably delete most of it, but occasionally there is something he will have to retrieve.

When he creates document, he usually names the document himself, and doesn't accept the suggested default name. He uses naming conventions for any files related to courses, using the course code, semester and year, as well as a short descriptor of the document. He keeps the master copies of most documents in Cecil¹. The documents he keeps locally are either ones that are currently being worked on, or are too sensitive to be placed in Cecil (e.g., tests and exams).

Brett occasionally has some confusion as to where the most recent version of a file is located, whether it is in his email or on the file system somewhere. Non-sensitive documents don't have this problem because their most recent version is always in Cecil. Normally, if Brett has to find a document, he will just browse through the folders where it might be until he recognises it. Only rarely will he search for a document, and that will usually be by file extension, or perhaps part of the name.

Brett's email is more organised than his file system. All emails related to any course are filed into a folder for that course. These course folders are arranged in a hierarchy of year and course. All other emails are filed into folders according to who the email was from, or particular issues e.g. (Jokes, Server Issues). Brett never searches for email. To find one, he just goes to the appropriate folder, sorts by from field (if necessary), and scans the subject until he find the message.

Summary of Interview with Participant C (Candice)

Candice is a Course Manager at Auckland University. She is responsible for administering a number of very large courses, and her duties include hiring, training and managing staff. She is in her 40s, and has been in the job full time for about 7 years. She believes that her documents and email are reasonably well organised, but that she never has time to get them properly organised.

Candice keeps all her documents in subdirectories of My Documents, and uses the tree view to navigate through them. She organises her folders around functional areas that she has to deal with. For

¹ Cecil is the name of the University of Auckland enterprise Learning Management System. Staff and students can access the system through a web browser. Staff with teaching responsibilities use it to communicate with students and to make documents and files available to their classes.

instance, she has a folder for final exams, a folder for data entry, a folder for mid-semester tests, a folder for coursebooks, a folder for labs, a folder for aegrotat paperwork and so on. Material relating to the current semester is stored directly in the top level of this folder. Other material that retains its currency, like templates, and instructions is often in permanent sub-folders of this folder. Material relating to previous semesters is stored in a folder called 'old' inside the top level folder. Within this, folders are named by year, and within that, by course code. This 'old' structure is duplicated in a number of different folders. Candice knows that the most recent file will always be the one in the top level. She sometimes uses dates in the filenames to keep track of evolving versions of a document.

Candice keeps a few documents that she refers to regularly on her Desktop. Many of the files stay on the desktop permanently, but the content changes over time, for instance a list of phone numbers, the office hours of tutors and coordinators, or the list of current textbooks. The location on the Desktop has some meaning, with items on the right being more active, and items on the left (particularly top left) being the least frequently used.

Every so often, Candice becomes annoyed by obsolete or poorly organised material, and might spend half an hour reorganising. This usually involves deleting files that are no longer needed. When Candice deletes, she usually goes into the Recycle Bin and empties it immediately.

Candice often opens recently used documents from the 'Recent' menu in Word. If she can't find a file, she will normally browse through the folders where it might be located. Very rarely she will search for it, and if she does, she will specify the location, and specify keywords from what she thought the title might be. Candice views her documents in the details view, usually sorted by name, but sometimes by file type or by date. She never sorts by size, and finds that size is meaningless to her.

Some of Candice's email folders are similar to her file system folders, with things like exams, tests etc. She also has a number of folders that map to people she deals with, divided in to Lecturers and Coordinators. She uses her inbox as a list of things to do, and tries to keep it to 30 items or less. Once things are dealt with, they are moved to the other folders. She also creates notes in her inbox to remind herself to follow up on things.

Every couple of weeks, Candice will need to return to a previously filed email. When she does, she goes to the folder it is likely to be in and browses to see if she can see the subject line. If she can't, she sorts by time to locate the message. She almost never searches for email. If an email has no value, she will permanently delete it. She routinely deletes Sent Items and Deleted Items. Candice also uses the Mark as Read function to ensure that all seen or useless emails don't show up as a bold folder name in the tree view.

The biggest problem Candice has is managing the volume of email she gets. She also laments the fact that she doesn't have time to sit down and organise her files and email 'properly'. The feature she appreciates most is the ability to use folders to group files into related topics.

Summary of Interview with Participant D (Damien)

Damien is a Course Manager at Auckland University. He is responsible for administering a number of new courses, and his duties include hiring, training and managing staff. He is in his 20s, and has been in the job full time for 2 years, and part time for 3 years before that. He is also currently completing his PhD. He believes that his documents and email are very well organised.

Damien keeps almost all his documents inside three folders on his Desktop. One folder contains all information relating to the courses he manages. Within this is a folder for each course. The documents relating to a current semester are in the top level folder of each of these course folders. Materials related to previous semesters are stored within folders by year and semester inside each course folder. Another top level folder on the desktop relates to administration, such as pay calculations. Within this, the folders are generally structured by employee. His third Desktop folder contains documents related to his PhD thesis. Within this folder, documents are grouped into readings, the thesis document, and presentations.

Apart from the three Desktop folders, Damien keeps one file on the desktop, a travel log, to remind him he needs to fill it in regularly. He has removed the system default icons that he doesn't use, such as My Network Places, leaving only My Computer and My Documents. There are also two backup icons that cannot be deleted, so he places these on the bottom right of the Desktop where they are least noticeable. He dislikes clutter on the desktop, and even goes as far as removing icons from the System Tray.

Damien uses My Documents to store documents that don't really have a place in his established hierarchies. Generally files in My Documents will either be deleted when no longer of use, or moved to a more permanent home. He also uses a folder on C drive to temporarily store some downloads, because it is easier to copy them to the correct location afterwards than navigate through the hierarchy in the Save As dialog box. When he deletes, he either deletes permanently, or immediately empties the Recycle Bin.

When Damien creates a new document related to one of his courses, he uses a naming scheme. He uses the course code and name, the document name and number, and then a version number. He often deletes older versions, and gives the final copy of the document a 'version final' designation.

Damien uses the tree view to navigate through his document hierarchies. He almost never searches for a document in his file system. He currently uses the details view, because his computer is new, but

normally uses the list view, especially when the folder contains more than one screenful of documents. The list view allows him to see more without scrolling.

One of Damien's biggest problems is documents that can be in multiple places. He puts it in the one that fits the most, but this can cause confusion when retrieving it. Also, the time taken to navigate down through the hierarchy to locate a file annoys him and he finds it a timewaster. However, the hierarchy is also his biggest benefit, as well as being able to change the sorting to locate things in different ways, such as by type.

Summary of Interview with Participant E (Edward)

Edward is a Lecturer at the University of Auckland. In addition to his teaching role, he is also responsible for the administration of a number of servers, and is also completing a PhD thesis. He is in his 50s, has been teaching for 3 years, and believes his documents and email are moderately well organised.

Edward keeps his documents in two different locations. Things that he is actively working on are kept in a folder called Working Articles on his Desktop. There are also a couple of other topic specific folders on his desktop that are also frequently used. Other documents are stored in subfolders of My Documents, depending on the topic they are related to. If Edward hasn't used a file on the desktop in a while, he will copy it into a folder under My Documents. If he hasn't use a folder in a long time, he will copy it into a special Edward Backup folder that he reserves for very old documents that are no longer needed.

Edward keeps many of his documents in more than one place. He keeps copies on his work desktop, his home desktop, his laptop, and on a USB drive. He changes files on whatever machine is currently in his vicinity, and then tries to copy the updates to the other machines as soon as he can to avoid synchronisation problems. However, he does encounter cases where he has two different copies of the same document, and needs to open both copies to check which is the current version.

Summary of Interview with Participant F (Frank)

Frank is a Lecturer at the University of Auckland. He is in his 50s, is responsible for teaching several papers each year, and is completing a PhD thesis. He has worked at the University of Auckland for 18 months, but has worked teaching similar courses for a number of years prior to that. He believes that his documents and email are pretty well organised.

Frank keeps multiple copies of his documents and folder structure. He has one copy on the hard drive of his laptop, one on the hard drive of his home computer, one copy on a network drive, and sometimes has a partial copy on a memory stick. He also has an almost identical structure in his email. He treats the network drive copy of the documents as the master copy.

The only difference between his master copy in the file system and the email, is that the email is ordered by course then semester, and the file system is ordered by semester and then course. Frank sees these as equivalent, and the structure below that (i.e., per course per semester) is identical.

Some of the other copies are not complete, missing either some documents, or entire folders. Frank uses a freeware utility to automatically synchronise between these multiple copies of the data. He occasionally has problems because he updated the wrong version of a copy, but generally, his utility keeps everything synchronised.

In some cases, he will not duplicate a folder in all drives, but will instead create a shortcut to the folder on another drive to minimise the possibility of having duplicates get out of sync.

In addition, Frank tries to standardise everything he does as much as possible. For every new course he teaches, he creates a folder in his email and a folder on his hard drive for that course. He uses VBA and batch files to automatically create the same set of subfolders under the course in both email and in the file system. In addition, he has some script files for automatically setting up some folders to classify his Program Files. He uses this classification both inside his Program Files to structure the installed programs themselves, as well as having a mirror structure on C drive which contains the install files (zips and exes) corresponding to the programs.

Frank intends to merge all his script and batch files together into a windows application that he can use to automatically roll out his structure on a new computer, and update the various versions to keep them synchronised as required.

Frank also uses naming conventions for files, often including the file path in which they are located in the name of the file itself. He does this because he foresees that the file may one day be moved or copied outside its folder context, and he wishes to preserve that contextual information with the file.

Frank keeps very little information on his Desktop. He uses it generally as a dumping location for documents from the internet or email attachments that he perceives will not need to be permanently stores. He saves email attachments to his Desktop rather than opening them in Outlook to ensure they are virus-checked.

Summary of Interview with Participant G (Gail)

Gail is a Lecturer and PhD candidate at the University of Auckland. She is responsible for teaching several papers each year. She has worked for the University of Auckland for 5 years, but has taught in other locations prior to that.

Gail keeps her files in her named Documents and Settings folder. The folders she creates are mixed in with system created folders. Although she didn't know who created those, she is careful not to

interfere with them. She is also careful never to change the names of files and folders created by other people and sent to her. She feels that as she doesn't own the file, she doesn't have the right to do this.

She navigates using the Windows Explorer tree view to locate her documents, and almost never uses search to find a file. She has a fairly shallow and broad hierarchy, and doesn't mind how many files are in a folder as long as she views them as logically related.

Gail feels fairly satisfied with Windows Explorer and feels that her mind works in a similar way to the software.

Summary of Interview with Participant H (Harriet)

Harriet is a contract Lecturer at the University of Auckland. She is in her 40s, is responsible for teaching several papers each semester, and is completing a PhD thesis. She has worked at the University of Auckland for 6 months, but has worked teaching similar courses for a number of years prior to that. She believes that her documents are very well organised.

Harriet has an extremely small file system on her laptop, kept in a folder next to My Documents called 'Harriet's Data'. She takes the laptop home with her, and it is her primary work environment. She regularly makes backups of her entire data folder onto re-writeable CDs, which she carries around with her, as well as having copies in multiple locations in case something happens.

Her primary segmentations are teaching and research, with approximately half a dozen subfolders in each. In her teaching, she keeps two previous semesters of courses in order to be able to give students past exams and assignment, and to enable her to reuse documents from previous semesters. However, after two semesters have passed, she will archive the entire course folder onto a CD and delete it from her hard drive. Because she only has one 'active' copy of her documents, she doesn't have versioning problems.

She very seldom creates new blank documents, preferring instead to find another document (either hers or someone else's) which has a similar structure, content or format, and will open that and use Save As to create a new document.

Harriet always knows where her files are and never needs to resort to search to find her own documents. She will occasionally search for email attachments that were accidentally saved in the wrong place, usually searching on a partial file name. She uses the list view most of the time, but very occasionally switches to details to check file sizes of files that are going to be distributed to students.

Harriet hates the Desktop, preferring to use the Office Toolbar shortcut to Windows Explorer to access her documents, and to use the Start Menu to switch between them. She can't see any reason to use the Desktop as it is always less efficient, and she prefers to accomplish things as efficiently and simply as possible.

Harriet tries to name folders and files consistently and clearly, so that someone else would easily be able to locate and identify documents in her file system. Similar documents are named in similar ways, and files that are going to be distributed to students often have the course code in the file name.

Summary of Interview with Participant I (Ina)

Ina is a Professor at the University of Auckland. She is in her 40s, is responsible for teaching several papers each semester, serves on a number of committees, and regularly writes journal and conference papers in collaboration with colleagues. She has worked at the University of Auckland for 4 years, but has worked teaching similar courses for a number of years prior to that. She believes that her documents vary from poorly organised to quite well organised. She says they deteriorate over time as she allows documents to just pile up in the top level of My Documents. Every six months or so she goes and cleans them up by moving them into the appropriate folder – the same thing occurs with her physical documents in her office on the same schedule.

Ina uses a laptop both at home and at work, and carries files between them on USB drive. While she is working on something that she knows will be an ongoing project, like a journal article, she will create a folder to store all the documents in. She will keep separate revisions as different files, giving each a date to distinguish them. When she is finished the article, she will generally keep the final submitted version and will delete the incomplete versions. These go into her Recycle Bin which she deletes about once a year.

She primarily accesses her documents through the applications she uses to edit them, using both the Recent Documents list on the file menu and the Open/Save dialog box. She sometimes uses the Recent Documents menu on the Start menu.

Ina always knows where her files are and never needs to resort to search to find her own documents. She uses the list view in the Office dialogs, and the large icon view when she opens My Computer. She accesses her documents by choosing My Computer on the Windows XP start menu and then selecting My Documents in the shortcuts displayed on the right hand pane.

Ina tries to name folders and files descriptively and clearly, in a way that is meaningful to her. She doesn't use many codes or naming conventions because she probably would forget what they meant.

Summary of Interview with Participant J (Jack)

Jack is a Professor at the University of Auckland. He is in his 60's and is responsible for teaching several papers each year, serves on a number of committees and is an active researcher. He has worked for the University of Auckland for 11 years, but has taught and researched in other locations prior to that.

Jack feels that he should keep his documents well organised, but generally finds that over time, he doesn't maintain his folders too well. He starts with good intentions, but ends up just shoving files in anywhere. To Jack, having well organised folders means that files are segregated by file type within broad topics.

Jack tends to use fairly shallow folder structures because otherwise he finds things get lost too easily. If a folder gets more than a couple of dozen files in it and he can see some obvious groupings (often file types), he will subdivide the folder.

Jack keeps all his documents in subdirectories of My Documents. He uses the Quick Start menu rather than Desktop shortcuts, so tends not to view the Desktop very often.

Jack uses the Windows Explorer tree view to browse through his documents, and will usually browse for a document if he has a reasonable idea where he is. Because he does not completely trust his own ability to retrieve files from his filing system, he uses a free Desktop Search tool to allow him to easily search for files he can't locate. He often remembers key words and phrases from inside the document he is looking for and the document search tool is useful in allowing him to use those memories to locate documents.

The Desktop Search tool has given him the freedom to relax his need to segregate by file type, and he has more recently started to create folders that assemble documents related to a particular folder or task, and mixes files of different types with the folder.

Jack tries to give his documents meaningful names, although doesn't have particular conventions for doing this. He does use version numbers for documents that are heavily revised.

4.2.2 Interview Analysis

The interviews were transcribed and coded using thematic analysis, using the QSR NVivo 2.0 qualitative analysis software package. Thematic analysis is a process of coding qualitative information (Boyatzis, 1998), that focus on identifying themes and patterns of behaviours (Aronson, 1994). The transcript is first examined for any major themes and patterns that emerged. This is followed by an iterative process of examining the transcript for any data that relate to the already identified patterns, while being alert to new patterns that emerge. The patterns can then be further classified into subthemes.

As much as possible, the theme and subtheme names were taken from the actual words and phrases used by the participants. *"If you don't want to learn from the data, but only aim to describe the patterns of people's responses to preconceived questions, you do not need to keep their words or details of their experiences. A survey would do. If you do want to learn from the actual accounts in your data, your aim is always to have ideas emerge from your working with data."* (Richards, 2005 p67).

The interviews were analysed and coded in the order they occurred. Initial coding was done at a very specific level, with coding for things such as an aversion to search, preferences for time based sort, or reluctance to delete. After the initial coding of each participant using the ideas and themes that naturally emerged, the themes were examined and related ideas grouped together as subthemes of a common theme (such as search, sort, delete behaviour). Additionally, the previous interviews were reviewed again to identify any instances of the newer themes. This was repeated through all the interviews until finally arriving on a set of stable themes and subthemes.

The stable set of themes and subthemes created after the analysis are shown in a conceptual model in **Figure 22** below. The following sections will discuss each theme and present the evidence for each theme that was derived from the interview analysis.

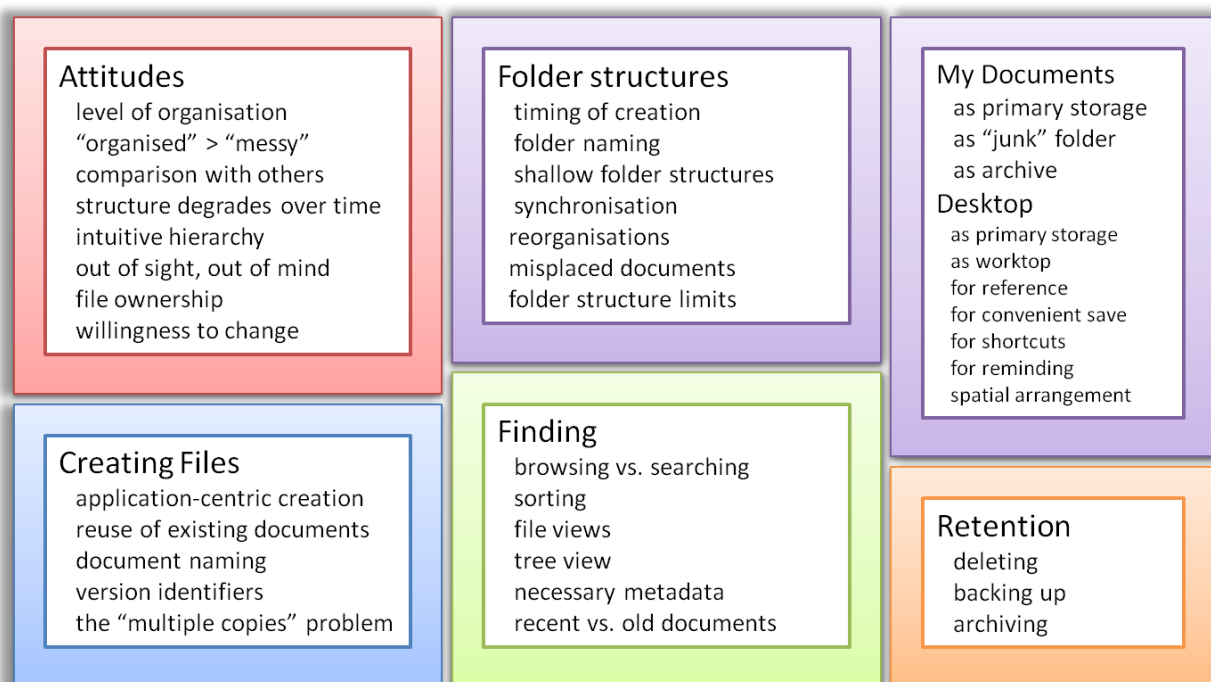


Figure 22: Conceptual model created from interview analysis

All quotes attributed are in the participants' own words. Quotes from participants who are second-language English speakers (such as Edward) may have unusual or inconsistent grammar. Any references to specific projects, courses, campuses, people or organisations that could lead to the participants being identified have been changed. Participants C and G (Candice and Gail) both requested that no direct quotes from their interviews to be included. Therefore, their findings are all paraphrased rather than quoted directly.

4.2.2.1 Attitudes

Alex describes himself as very disorganised, and later remarked "*I fully admit that I'm pretty hopeless at managing files.*" Brett also considers himself pretty disorganised.

Candice rates herself as average. While she doesn't have any particular difficulties, she does wish she had time to clean up her files, and worries that sometimes she creates folders for documents when there is probably already an appropriate folder somewhere if only she could remember where. Frank also gives himself an average rating, saying *"I try to organise my files, it is not easy, it is easy to say, but difficult to do, how to well organise the file structure on my computer."* Gail considers herself averagely organised, and although she sometimes feels that she should be more organised, it doesn't really bother her. Jack gives himself an average rating, saying that he was initially *"very conscious and careful about producing clear directory structures and then latterly I just shove files anywhere,"* although his search tool means that this lack of organisation doesn't impede him finding documents quickly

Ina rates herself as above average in organisation, but says her rating varies with time. After a reorganisation, she would rate herself as extremely well organised, but she lets it drop down to a fairly disorganised state and then thinks 'oh I can't stand this anymore' and reorganises. She notes that this is *"my perception of what's organised in terms of a system that works for me."*

Damien rates himself as very well organised, saying *"I'm not anal about it but it's fairly organised. I know where everything is usually."* Harriet also rates herself highly, saying, *"Most of the time I actually know what my files are. It's scary, but I can usually find what I want immediately because I take care in using consistent naming."*

While this is a very subjective assessment, after taking a tour of the participant's files and examining their file system snapshots, the researcher believes all the participant's self-assessments are accurate.

Being "organised" is a desirable state

Many of the participants' responses indicated that they felt that being organised (whatever that means to them) was a desirable state. Those who consider themselves organised expressed pride in their file structures, for instance, Ina spoke with pride of her colleagues being *"surprised that they can come into my office and they can ask for an article and I will know where it is,"* and that *"people will tend to say Ina will probably have it",* adding *"thank God for the power of computers."* Harriet also mentioned with pride that she wanted her file system to be easily understandable to anyone who had to look at it, *"without having to wade their way through some bizarre coding system."*

Alex described himself as *"pretty hopeless at managing files,"* but mentioned that he was much better in the past: *"you should have seen me when I was a student I had much better folders. I had a student query a mark from their 1999 tutorial, sorry, 1999 assignment, when it was about 2002, so I had to go back. Luckily I'd kept all the archives and stuff in zip files."*

When asked if he had any patterns for naming files, Edward said that he didn't, but that he should have. Candice says she feels she should take time to reorganise her files, but she doesn't. At several

points during the interview, she indicated various files and folders and said that she should go through them, but hadn't had time.

Jack noted that he feels it is more important to organise some of his documents than others: *"I'm probably more organised when it comes to lecture materials, particularly PowerPoint files, than anything else. I take that very seriously."* At the end of the interview, Jack said that he found the interview very interesting, saying: *"it makes me think about, it makes me ask myself the question, should I be relying on these crutches like Desktop search and so on or should I be more organised, and I think my answer to that is, it's a good crutch! I'm gonna stick with it because I don't have to worry."*

Being "messy" is an undesirable state

Several participants referred to their documents as being messy, or a mess. Edward, upon opening a folder said *"I told you that sometimes a mess,"* and Frank similarly said *"it's a bit of a mess at the moment actually."* Candice also described one particular folder as a mess, and Jack described his folder structure as *"messy really,"* saying that it was *"a jumble."*

Comparison with other people

Several of the participants compared themselves to other people, or mentioned how other people they knew organised their documents. For instance, Ina mentioned that her husband tells her she should use a search tool to find her documents, and remarks that *"the way my husband files, he's always searching."*

Candice asked about how other people organise their documents so she could see if she did anything similar. Jack also asked about how other people organise. When talking about his transition from file-type based folders to topic based folders, he asked, *"how do most people organise themselves? One or the other I guess, but not both?"* and later remarked that his folder structure *"could be a lot better. I'm sure that most people are better."* When discussing the Desktop, he commented on a colleague whose Desktop he's seen, saying *"she needed about five screens I think to scroll down,"* and then remarking that he found the interview *"very interesting, I'm glad I signed up to it. You can learn a bit about yourself and get some comparisons with others. The diversity of approaches is I wouldn't say astounding, but it's quite wide, isn't it?"*

Document structure degrades over time

Jack observes: *"when I started with a new system I was very conscious and careful about producing clear directory structures and then latterly I just shove files anywhere,"* adding that *"my folders got rather messed up, and I can't be bothered to get back and sort them out."* Ina notes that she lets her structure degrade until they are in a disorganised state and then cleans everything up back to very organised again. Alex observes that in his previous role as tutor, he was much more organised but that now he no longer bothers.

In contrast Edward says he hasn't changed anything about his folder organisation, as his document management practices have now become habit. Damien also notes he hasn't changed anything except moving his primary storage location from C drive to the Desktop where it is easier to access.

Intuitive hierarchy

Several people described the folder hierarchy as being intuitive. Frank said that folder structures are *"second nature, and I probably don't feel things that someone who is new to them would find puzzling and annoying. I'm reasonably happy with this hierarchical tree structure of Windows Explorer."* Gail noted that she felt quite satisfied with Windows Explorer, saying that her mind works in a similar way. Jack also mentioned finding Windows Explorer easy to use: *"I find Explorer pretty good, it's been around quite some time and it's just, it's so facile, so easy,"* adding that, *"it seems to fit in with my mindset."* Ina says that she's *"always been a PC user, I haven't come across from Mac or Apple so I tend to think quite intuitively like these guys do."*

Out of sight, out of mind

Damien explained that he likes to have things placed into folders, saying *"I don't like having clutter, I can't stand clutter on my Desktop."* He notes that he's removed most of the system generated icons from his Desktop, and even on the Taskbar notification area, he has removed most of the icons, saying he prefers to have things hidden until he needs them.

Ina and Brett both mention the idea of putting away old documents, with Brett talking about moving everything and starting again with *"a clean slate,"* and Ina saying *"I'm the kind of person that if it's not there, I don't even look at it."*

By contrast, Alex prefers to have everything visible, saying he doesn't like moving things from the Desktop into folders because *"my fear would immediately be that now I've moved them off my Desktop, now I won't know where they are."*

File ownership

Gail has a very strong conception of file ownership. She related a past experience where she was assigned a computer that previously belonged to someone else and still had the previous user's files and folders on it. She was very careful not to move, rename or delete or in any way interfere with that person's files. Windows XP automatically creates a number of folders for each user, including folders for Application Data, Local Settings, Templates, and known network computers and printers. Several times Gail referred to these as being someone else's folders that she didn't touch, seemingly not knowing they were system created. She mentioned being very careful not to use them or touch them. She also feels that she doesn't have the right to rename files that she didn't create herself. Any files she saves into her folders that were emailed to her or that she downloaded from the web always keep their original

names. She explains that even though they are on her computer in her folder structure, she doesn't feel she has the right to rename them because she didn't name them and she doesn't own them.

Harriet also mentioned file ownership, saying that she doesn't like other people editing files she owns. Rather than collaborate with someone by sending an electronic document back and forth in multiple versions, she'll have the other person print out the document. She explains that her collaborator *"usually prints them out in hard copy and writes all over them in red and then I can choose what to accept and what to reject."*

Willingness to change

During the interview, Edward said he'd like to improve his document management, saying he would like to change *"if you could tell me after this research what is a good way or better way to organise files, that means is easy to name it and easy to retrieve it."* Damien noted that he'd like *"a better way"* of dealing with his archive folders, but was otherwise satisfied. Most others indicated they were relatively satisfied with their document management practices, although Ina commented that she doesn't think she uses the system as well as she could. Gail mentioned that she was quite annoyed at having to open My Computer and then drill down through folders from My Computer to C drive and down to her document folders every single time she opened a document. At the end of the interview, the interviewer mentioned that it was possible for her to create a shortcut to her documents folder and put it on the Desktop for fast access with a single-double click. She politely said thank you, but was not interested in creating a shortcut, explaining that she was used to doing things a certain way and wanted to stick to the methods she was used to as she knew they were reliable.

4.2.2.2 Creating Documents

Application-centric document creation

All the participants create new files by opening the application they wish to use and then saving their document from there. This has significance because it means that in many cases, during the steps of creating and acquiring documents, the document collection is seen through the lens of the File Save Dialog box of whatever application they are using. While there is a standard system Save Dialog box, not all applications use it, and thus the view of the documents could be significantly different during the creation/acquisition stage than it is during the organising/retrieval stage.

Alex and Frank report that in some circumstances they create the files in Windows Explorer instead. Alex does this with documents he intends to permanently live in My Documents, as opposed to documents destined for his Desktop which he creates through the application. Frank creates text files by right clicking in the folder and creating a new text file, giving it a name and then opening it to add the content.

When Alex creates his documents through a Microsoft Office application, he always types in a *“very detailed and long title”* into his document and then saves it, letting Microsoft Office pick up that document title as the file name.

Ina also creates new documents by opening a new blank document in the appropriate application and begins typing into her document before she saves it but she doesn't allow Microsoft Office to use the title as the file name, instead replacing it with something that is more meaningful to her. Brett, Candice and Edward likewise report usually saving and naming the document after working on the document for a while. Brett noted that if the document was expected to have no lasting value, he would use the default name rather than assigning one.

Damien makes a point of saving into the proper folder with the correct name before typing anything. When Frank creates documents via an Office application, he also tries to save it before adding any content, often also adding metadata such as title and author to the file via the properties dialog. Jack also begins writing before he saves

Harriet seldom creates new documents from scratch, but instead usually opens a similar document and then uses Save As to create a copy. The document she bases her document from is one that is similar either in content, structure or style and formatting. Gail also tries to do this as much as she can.

Reuse of existing documents

Several participants specifically mentioned that they try to reuse existing documents as much as they can. Damien mentioned at the start of each semester going back to teaching material from the previous semester to reuse it.

Harriet tries to make extensive reuse of existing documents, saying she creates brand new documents *“as seldom as possible.”* She says, *“generally if I've got a document, there's a reason I needed it, and again often when you create a document, you may not have done that thing before but you may have a document which is, 'oh I put a proposal together for such and such, I'll use that proposal document', it just saves rethinking about things.”*

Ina also mentions reuse:

“I can go back to old lecture notes and see what's there that I want to bring forward. I can go back to writing an assessment of somebody for promotion or a referees report for somebody. There's kind of a skill and an art to that so rather than starting all over again, I'll go back and see basically what I did the first time and kind of do a little cut and paste and so on.”

Naming documents

For the most part, the participants did not have any kind of predefined naming scheme or pattern that they used. Brett, Damien and Frank all reported that for course related documents they might try to stick to a pattern, but this did not appear to be a firm rule with any of them. Brett says of naming

schemes *"I have tried it various times only when they're directly related to a document for a course."* Damien says *"I've started to use those. I didn't really use them before. It was more just assignment 1 and the folder was the convention, but now I tend to think ahead that I'm going to put them up on Cecil and stuff and it just makes it easier for me, so now I do."* Frank says he uses a pattern *"more often than not"* with course related documents.

Generally, the naming scheme involves the course name, code or number, possibly a semester and year indicator and a description of the document. Brett's convention is *"course code, semester, campus, then normally followed by an underscore then the actual name of the topic"*, with Frank making sure to *"include the name of the course and probably a semester indicator as part of the document's name."* Damien says, *"my conventions for like assignments and things like that would be course name, course code, description, number and version."* Gail had a similar pattern that she used sometimes but didn't consider it a convention that she tried to stick to.

Frank observes that there is some information duplication involved in this: *"Usually it's going to be saved in the folder that's got the same name, but of course documents do go adrift, or find themselves in other contexts, so I'm looking ahead to those times when the document has been copied to or moved to a different context, I like to keep the course related information as part of the title."* Damien also noted that the information about the course and semester is also contained in his folder names, but that he duplicates it because he knows the files will eventually appear in a learning management system where they will no longer have their folder context.

Dates frequently figured in these naming schemes, most commonly being either a year or a semester identifier (e.g. '2004 S2 Report'). Alex and Edward referred to sometimes using dates in their file names, and took care to use a reverse date format (e.g. '2003-12-05') in order to ensure that documents names with a date appeared in date order when the name was sorted alphabetically. Harriet also noted that she names things in order to force a particular sort order.²

The other participants didn't use any particular naming conventions. Alex said explicitly that *"I very seldom use naming conventions. I normally give them, the documents, a descriptive title"*, later describing his file names as *"very descriptive and very verbose"* and saying that *"they don't contain any fancy coding or anything cause you don't put codes in your title of your document."* He likes to have the title of the document to be the same as the file name, which often results in his files having very long names. However, he notes that *"certain projects tend to have particular names that apply, for example,*

² With a standard alphanumeric sort, Lecture10 will appear between Lecture1 and Lecture2 because '10' comes before '2' in an alphanumeric sort. In order to force the items into numeric order, the items with single digits have to be named Lecture01 and Lecture02.

Project X project, all my files that are to do with Project X have Project X in the name of the file, just by definition, not through any conscious effort though."

Harriet doesn't use any naming codes:

"I sort of work on the basis that someone should be able to come into my computer and find a file without having to wade their way through some bizarre coding system. I mean if I dropped dead tomorrow, someone could theoretically come in and find all of my teaching stuff from the most recent to archived stuff and actually be able to use it without stressing over 'what's this', 'what does this one mean'?"

Ina also eschews codes: *"I don't use numbers, I don't use acronyms, I don't use hidden codes. I can barely remember my PIN number so I need to have stuff quite literate."* She tries to make it something intuitive so that she can look at it and figure out what it is: *"something that is meaningful to me."* Like Alex, she says her filenames are sometimes quite long and reasonably detailed.

By contrast, while Jack tries to make sure his file names are descriptive and relevant to the topic, he tends to give fairly short titles, because in many views, the full filename is not visible.

Edward, Frank, and Harriet mentioned renaming things they received from other sources. Edward saves a lot of HTML files and PDF files he encounters on the web. He says he often renames them because the HTML file names can become extremely long, potentially causing problems with backups. Harriet will normally rename articles she's downloaded with the same form a reference would take (author and year). Frank explains his reasoning for renaming files: *"Other people are often not as pedantic as me with the naming of their files, and files come in attached to email with rather, well, a flyer for a seminar will come in as seminar.doc. So if the file has got a name that's just one of those generic blah names, and I intend to make a permanent copy of it for myself, I will rename it to something that conforms to my systematisation."*

By contrast, Jack says he usually just keeps the original name. Gail never renames anything she downloads or receives from elsewhere. Since she doesn't own the file, she doesn't feel she has the right to rename it.

Several participants mentioned having temporary files. Alex pointed to another file called 'foo.csv', saying *"by its name you can tell that it's a very temporary file, I just created it one day because I was, analysing some data". "That file is no longer needed by me and I would be unlikely to put it in my My Documents, because My Documents is for files that actually have a value and that I might want to revisit later on. So foo is more likely to end up in one of my little archive things but it's also equally likely to be deleted straight off the system."*

Alex also indicated another temporary file he'd created, saying *"It's not a document per se it's more like a file I generated but never want to revisit again so it ends up in my little cleanup folder."*

Brett spoke of not bothering to give proper names to files which he knows are temporary: *"If it's something that I know is, like, immediately temporarily, and will be deleted in like half an hour then I won't bother [naming it]."*

Damien also noticed and found some temporary files while he was showing his documents: *"Most of this can actually be deleted now that I've looked at it. [Selects about 6 documents] Cause it's just temporary crap. I want to keep that ... [Selects a document then deselects it] I want to keep those ... [Deletes the selected documents]"*

Version identifiers

Almost all the participants kept multiple versions of documents in separate files and added identifiers to the file name in order to distinguish between versions of the document, and in particular, so they would know which file represented the most recent version of the document.

The most common method of version identifiers was to append a **number** to the file name, as Brett explains: *"Normally, they'll share the same name followed by an underscore and a sequential number, 01, 02 and so on."* Brett has many versions of some of his documents: *"something like the thesis, that went up into like a couple of hundred versions of that. Coursebooks generally have a half dozen versions. Exams normally have a half, maybe a dozen versions."* Because of his sequential numbering, he normally doesn't have any difficulty in figuring out which version is the most recent.

Damien, Edward, Gail and Jack all reported using the same system. Damien also noted that sometimes the final version in the numeric sequence would be given a different version identifier to indicate it was the final product: *"I'll have a final version which is the production one. And I call that final version as well so that I don't get mixed up."*

Edward also noted that he applied version numbers after the fact if he discovered two copies of the same file in different locations and wasn't sure whether they were the same or not: *"Normally I just check the date if they are the same or not, or the file size, they are the same or not. But very frequent they are different, by very minor changes of the date or the file size. And sometimes I keep them all there, and I name it 1, yes for example myfile1 and myfile0 depending on the date. So the zero one is the earliest date one, because at that moment, when I do the file reorganisation, I don't want to take too much time to look into detail, but I don't want to destroy the other one too. So I don't want to use different name, so I just use a little bit different name, but I still put them there, just in case I do need them at the same time."*

Adding a **date** to the version file names was also very common, with Brett, Candice and Ina all reporting doing this with some of their documents. Ina describes her process: *"So this is a paper that I've just been working on recently. It's a revision so I've got them all dated. [Topic] for this journal and a*

certain date. So we've got February, 14, 16, 21 March and then this is the response letter that I've been working on. I will get rid of all the old stuff once the paper is in print. But until it's in print, it all stays."

Ina and Candice noted that they often have both people's **names and dates** in documents that they are collaboratively working on. Ina describes this: "We get into a rhythm, if I use Alice for example, she has her own way of labelling, as I have my own way of labelling, so I might keep portions of what she has and then I'll have 'cb Ina' meaning changes by Ina and then give the date. So she'll recognise her file but it's been changed by me." When working with her colleagues, they all usually add their initials to the file name "so we all know who we're talking about when we're passing stuff around." Different conventions apply to different collaborators: "I have a colleague in Scotland who we don't use our names anymore, we're ping and pong. Because this paper we were working on was just going ping pong ping pong and we now, she's ping and I'm pong and that's the way we do everything."

Frank and Alex said they didn't keep versions. Alex said he relied on the track changes feature within Word for his version management.

The "multiple copies" problem

Four participants reporting have problems resulting from multiple copies of the same document (excluding backups).

Alex reports sometimes ending up with multiple copies of the same file in different locations: "what happens is I generate a document, it ends up in one of my cleanup folders, and I can't find it, or I forget that I had it there. I generate it again by whatever means, and that ends up in another cleanup folder, so I've got two copies of the same document lying around." He adds "it is also likely that a file on my Desktop is in My Documents as well, but they may be out of sync, so I may have created it in My Documents, copied it over to my Desktop so I can work on it and not synched back to My Documents." He says he rarely detects that sort of thing, and if he notices in a search that he has multiple copies, he will compare dates and use the most recent document.

Damien ran into trouble after trying to integrate files on his USB flash drive with his Desktop: "I keep my main copy on here now [USB drive], but then I copy across, but then I changed stuff and then I forgot to delete this one first [the copy on the hard drive], so it's double copied things because I changed the file names, which is bloody annoying."

Edward has fairly often ended up with multiple copies of the same file. He says "just in case that I lose some files in somewhere, most of time I keep three copies, or even four copies. For example I have a memory sticker [USB drive], I keep some file there and also here [Desktop], and also my laptop and my home computer. Sometimes it's a problem because there are too many different kind of versions, so different kind of copies." He tries to avoid problems by trying to immediately synchronise any files he changes, but he doesn't always remember to do so.

Frank has a special tool to keep all his document collections synchronised, which he runs periodically to make sure all his files are updated with the latest version. It notifies him if there are any problems synchronising and gives him choices about which files to keep and which to overwrite. He does very occasionally run into problems after having independently edited more than one copy of the same file: *“when the situation arises, more likely with a current document where I have inadvertently edited the wrong disk’s version of it. And then a day later, I can’t remember quite which of the versions I was editing. I may have to open up all of the versions to find out where they are and open them all up to make sure which is the one that I most recently changed.”*

4.2.2.3 Folder structures

Timing of creation - in advance, just in time, cleanup

Folders are created for a number of different reasons. They can be created before there are files to be placed within them, created ad hoc to contain files needing to be saved, or created in order to clean up and move existing documents. Most participants reported using multiple folder creation tactics depending on the circumstances.

‘In advance’ creation of folders refers to creating folders before having any content to put in them, although with the anticipation that there will be content for the folders in the future. Damien has a number of empty folders that are waiting for content to fill them: *“Projects [clicks on it, it is empty], oh, that was the project papers. Oh, I don’t think I ever used that but I created it in case I wanted to, but yeah, ‘cause we do do project stuff, but I don’t think ... yeah, I don’t think I’ve ever actually used that folder, but I’ll leave it there because it makes sense.”*

Candice also reported setting up folders for some new courses she was taking responsibility for, in anticipation of having relevant documents later on. As it turned out, she didn’t actually use them, instead documents relating to those courses were integrated into her existing document structure for her other courses.

One particular kind of advance folder creation is the creation of entire folder structures that duplicate an existing folder structure. Frank says, *“I generally have the same structure of subfolders per semester: admin, awards, a backup folder, a few miscellaneous things there. This one is where all the assessment material goes, labs, lectures, material related to [another] campus, material related to the [other] campus.”* Candice and Brett also refer to creating folders for a new year, with the same set of folders as the previous year.

Frank goes to the lengths of creating batch files and scripts in order to automate the creation of folders so that he can reuse his structures:

“I intend to expand that idea to other areas of my filing system, to standardise structures as far as possible. When I get a new computer, at the moment, that structure I’ve just

shown you, I've got a batch file that will create that structure for me in the current folder wherever I am, but I intend to translate that into a little Windows scripting host .vbs file, and give more flexibility and having done that, have a templated .vbs file utility that I can use to set up similar structure, standardise structures from machine to machine. It seems to me to be the way to go to keep track of material. If you move to a new computer and just ad hoc set up a file structure, then it becomes harder and harder over time to remember precisely where you've stored which sorts of material."

As well as scripts for creating his standard document folder structure on any computer, he has scripts for creating the same set of folders in his email client.

'Just in time' creation of folders refers to creating folders as the need arises to save a file which does not currently have any location. Damien demonstrated saving an email attachment, saying *"first thing I'll do is I'll create a place for it,"* which he did using the New Folder option in the File Save Dialog box. If Candice goes to save a file and can't figure out an appropriate folder for it in her existing folder structure, she'll create a new folder at that time to hold it. She doesn't need to do this often anymore since she finds her current folder structure normally accommodates most of the files she encounters in her routine work. Frank mentioned creating a folder to hold some information he found during a search, anticipating that as he continued searching he would find more items to add to the folder. Harriet also described a research project folder that currently contained a couple of items, saying *"at the moment there's only a couple in there, as that grows, that'll get subfolders for each of the companies I'm involved with."*

Rather than use a 'just in time' creation strategy, some participants instead would save the file in a temporary location (such as the Desktop or the top level of My Documents) and then later 'clean up' by moving it into a permanent location (and creating a folder for it if one didn't already exist). For instance, Ina tends to do it this way: *"I'm quite ad hoc to begin with. I often just save it into the open My Documents and then I figure out where it all goes later."*

Cleanup creation of folders refers to creating folders in order to group and organise files that already exist. Alex mentions this on a couple of occasions: *"my Desktop fills up to a point where it's unusable, and then I just pull out all the unnecessary stuff and pop it into a cleanup folder",* and *"I once created an Project X folder for the Project X things."* Brett follows a very similar approach: *"Periodically these files get dumped into folders on the Desktop. Periodically those folders get dumped into, I guess other folders either in My Documents or at times onto the C drive. Like I'll create a folder which will have stuff on the first of the first two thousand and something [01012004]. Which means I'll pretty much start again with a clean slate."* When Candice thinks her folders are getting to full with old information, she'll create a folder called 'old' and move all the older material into it, leaving only the things she regards as current.

Harriet also mentioned creating folders to clean up: *“I shifted all the old stuff into a subfolder, so if I’m finding a folder is getting cumbersome, I’ll simply create a subfolder and shift everything in it, or clean it out.”*

Folder naming

Folders are named in a number of different ways. Participants’ self-reported folder naming patterns are analysed here, and their actual folder structures are analysed in **Section 4.3.5**.

Given that the participants were all from a university context, it is not surprising that most participants reported having folders named for **courses** they were involved with. At the University of Auckland, each course has an alphabetic department code and a 3 digit course number as well as a course name. While a couple of participants used only the 3 digit codes to identify their courses (Candice and Harriet), most used a combination of the course code and number.

Time is an extremely common element in folder names. Brett has ‘history folders’ into which he regularly dumps file and folders he no longer uses. These are named with dates (e.g. ‘01012004’ and ‘070502’). In addition, his teaching folder is segmented into years, with folders named 2004, 2003, 2002 etc under which he created individual course folders. Jack changed his habits recently: *“I had everything in Lectures and then I realised of course that last year’s set of lectures is changing for this year so I’ve now got them under the year.”*

Damien and Gail also keep folders named by year into which they archive their current folders at the end of each year. Ina also has year folders but she uses the current year folder as an active working location, rather than archiving into year folders at the end of each year as Damien and Gail do.

Candice uses this same approach in some parts of her folder structure. In other parts, she doesn’t have separate folders for each year, but has a single ‘old’ folder into which she places all previous year’s documents, keeping only the current material in the top level folder. Harriet also has a hybrid approach, with folders named with year and semester in her teaching folder, and with folders such as ‘Old Admin’ in other areas. Frank uses the year and semester approach, having folders named ‘2004 Semester 1’ and ‘2004 Semester 2’.

Many people reported creating folders for particular **topics** they needed to store information about. Brett was pointing out some of his folders, saying *“most of these are related by topic as opposed to file type. That was a reasonably recent sort of thing.”* Damien has a folder called Admin containing all material related to administration of his courses, as well as having folders related to Wedding and Property *“because I’m getting married and I’m looking for a house.”* Edward mentioned doing some research work on Logistics, and creating a folder to contain all the material related to that topic. All the other participants also had examples of folders they created because all the files contained within were related to the same topic.

Many of the folders created group files according to the **genre** of the document. For instance Damien has folders for timesheets, for books and for coursebooks. Frank, Brett, Gail and Harriet all have folders within their courses for assessments, labs and lectures. Candice has separate folders for exams, tests, reports and 'forms and letters'. Harriet notes that within her PhD project she separates documents into separate folders containing presentations, conference papers, admin forms and thesis chapters.

Brett and Jack both reported creating folders for particular **file types**. Jack says *"I wouldn't put a Word document into a PDF folder. I've got PDF folders I've got Word Doc folders, Excel folders and so on."* He says that recently he has started creating folders for projects in which he mixes file types. Previously he kept them separately and notes that he still does so at home. Brett says that *"some of the folders are set up for "Word Docs", "Excel Docs", "Access Docs". But generally if that happens then, I mean like, all the Access stuff is related to the same topic anyway."*

Many participants reported naming folders for particular **projects or tasks** they are involved in. For instance, Edward, Damien and Harriet are working towards their PhDs and have folders called PhD which stores material related to their PhD research projects. Ina creates folders for each conference paper submission or journal article she writes, as well as having folders for committees she serves on or conferences she is organising. Candice, Harriet and Gail also have folders for committees they are on and organisations they are members of for which they need to do some work.

A few participants reported naming folders for **people** they worked with. Ina and Brett both have folders for the people they collaborate closely with. Harriet has a folder named for her husband as she typed some reports for him on her laptop while away on a trip.

The participants combine these elements into a subfolder structure in many different ways. For instance, Gail says that she structures her folders by course and then by genre and then by year. Frank starts with course, then subdivides into semester and finally by genre, whereas Candice organises first by genre and then by year and then by course. Brett and Ina both organise first by year, then course then genre. More information about how the participants actually structure their folders based on their file system snapshots is in **Section 4.3**.

A number of folder names do not fit into these categories. For instance, Damien has a folder called 'Temp', which he uses as a staging area for files he downloads from Cecil. Jack has a 'Miscellaneous' folder that is *"just a dumping ground, it's just got all sorts of things that I can't find anywhere else to put them but don't yet want to get rid of them."* Similarly, Harriet has a folder called 'Rubbish' containing *"things I'm not quite prepared to get rid of like early drafts and things I don't use now, but I'm not quite prepared to [delete]."*

Edward also says that sometimes he will create a folder called 'temp' on his Desktop and move items into it temporarily until he can move them to a more permanent location. Alex has multiple folders on his desktop called 'cleanup1', 'cleanup2' etc in which he dumps items from the desktop. There is a temporal element to these folders, since the folder number indicates which is more recent.

Shallow folder structures

Four participants commented spontaneously that they tended not to have very deep folder structures.

Alex says *"I tend to have very flat structures"* and *"I tend to have very very shallow structures."* Jack says that *"I tend not to like very deep folder substructures because I just get lost, I just can't remember where things are. I always will have a folder and a subfolder and a main folder and subfolders, but you'll notice that there aren't too many folders beneath that."* Ina also observed that she tends not to have subfolders inside her main set of folders. Harriet noted that her current document structure was very small: *"it's a nice small folder situation. I have a much more complex one from my old job, which had 8 years of folder in it."*

Synchronisation between multiple structures

Brett commented that *"having multiple file structures and multiple locations for the same material ... leads to a bit of hassles on occasion."* However, he doesn't do anything specific to deal with the issue. Frank, however uses some batch files to ensure that all his locations contain the same folder structure, and uses a third party synchronisation tool to ensure that they always all have the latest version of his files. He keeps the same set of documents on his laptop hard drive, a network drive, his home computer and also uses the same folder structure in his email. He finds that this makes it much easier to locate his documents, and he rarely needs to resort to search: *"by having a standardised structure, more often than not the document is where I think it should be."* He also finds that because all his locations contain the same set of documents, when he does search, he only needs to search through one collection as they are all mirrors of each other.

Reorganisation

Many participants spoke of cleaning up, organising or reorganising their files.

Edward mentions reorganising every six months or so, quite often at the end of semesters. He goes through folders, especially ones designated as temporary and moves the documents to a more permanent location. Ina also has six monthly reorganisations, coinciding with when she reorganises her office. She says *"I look at all the stuff that hasn't been saved into a folder, and I figure out if I need to delete or move it or whatever."*

Harriet once a year moves things into archive folders, and Gail has a similar schedule and process. Brett says he doesn't really reorganise as such, but that *"I move, only when I know that it's not going to be used again, so probably not very often. Generally it gets dumped and moved into a storage folder."*

Damien changed the location of his primary documents from the C drive to the Desktop. He used to prefer having a completely clean Desktop, but now allows three folders to stay on the Desktop because it makes accessing them so much more efficient. He does try to periodically clean up his My Documents folder, which he uses as a *"junk folder"*: *"not very often, I'd say, it's just once I realise that I need to clean it up, I will, and usually, I don't realise it that often, it's just cause, I don't really work with it, I just use it as a dumping ground for stuff that I might need, but not immediately. And now that I'm looking at it I want to clean it, you know."* Damien also mentions archiving at the end of each semester, deleting or moving items that he knows he'll no longer need.

Alex has a system of cleanup folders on his Desktop into which he periodically moves items from his Desktop. He says these are his attempts at reorganising and that he only does it when he has to. He does mention past attempts at organising: *"I had a current folder [opens it, from inside My Documents]. I was trying to have some sort of current folder, which had the current files in it, and all these things would be archived files so I was trying to be clever like that. Didn't work because what happened was that the current files, after about half a day, become not so current and it loses its appeal."* On another occasion he tried creating a folder to contain all the documents regarding a certain project he was working on, however that didn't work because he still kept ending up working on other documents related to the project on his Desktop. He has now largely abandoned the idea of trying to organise his documents. He says *"the only reason I would try organise stuff is to remove clutter so that I can do proper searching."*

Frank doesn't explicitly reorganise, saying his folder structure is a *"work in progress"*. Over the past 12 months he's been establishing parallel structures in his email and multiple computers and disciplining himself to use the same structure for each course.

Candice says she wishes she could reorganise her documents but she never has the time to do so. She sometime starts and gets partway through but always something else comes up that prevents her from finishing. She describes it as a constant guilt that she knows she should do something but never does. She does say that maybe once a month or so something will annoy her about a certain folder and she'll try to delete old items or move things to better locations, however she usually never finishes.

Jack no longer reorganises his files. He says that *"when I started with a new system I was very conscious and careful about producing clear directory structures and then latterly I just shove files anywhere."* Now that he uses Copernic Desktop as his search tool he doesn't bother to reorganise

because he knows *"I can just pull them back by searching for them without having to know where they are."*

Brett, Ina and Jack all mentioned the ease of dragging and dropping information around. Ina explained that she normally just drags and drops things around during a cleanup. Alex mentioned that when he does move files, he tends to copy and paste documents around rather than drag and drop much, fitting in with his method of navigating with the keyboard and not using the tree.

Misplaced documents: "that shouldn't be there"

Many of the participants noticed misplaced documents during the interview, with several of them actually moving or deleting items during the interview.

Candice noted a file on the desktop that had been there for a while, saying she doesn't really need it there and should get rid of them, and Gail commented that she needed to move some of her documents into year-based archive folders.

While looking through his primary documents collection, Damien remarked *"some of these needed to be cleaned up now that I look at them."* He didn't touch them, but later, while looking at his My Documents folder, he did some cleaning up:

"Most of this can actually be deleted now that I've looked at it. [Selects about 6 documents] Cause it's just temporary crap. I want to keep that ... [Selects a document then deselects it] I want to keep those ... [Deletes the selected documents] Yeah, so, not very often, I'd say, it's just once I realise that I need to clean it up, I will, and usually, I don't realise it that often, it's just cause, I don't really work with it, I just use it as a dumping ground for stuff that I might need, but not immediately. And now that I'm looking at it I want to clean it, you know."

Later in a different folder he notices two folders that have related content and observes:

"Department, this, ok, this one, and the Admin are probably one in the same [browses the Department folder, which has 5 items] cause, hmm ... yeah, I should actually get rid of the Department one ... copy all these over [selects and drags the messages into the Admin folder and deletes the Department folder]. Cause this is where I just keep crap stuff, like from [the Departmental Administrator] and ... non-academic material."

After Ina said that she doesn't keep documents on her Desktop, she noted that she had one folder of items from one of her master's students who had copied it from a USB memory stick onto the desktop. She says *"that's a mistake, that shouldn't be there."* When looking at the files in the top level of her My Documents folder she noted *"I thought it was fairly clean but actually it's amazing how much junk fills up that quickly. So that's stuff's all on file. Isn't that interesting. That's about 37 papers. 36."*

Several times during the interview Jack noted items that were out of place: *"That really shouldn't be there, it's probably nothing anyhow, I don't know what that is. So you can see I'm not that well organised."* Later, he noted a folder was getting a bit full and talked about some of the contents, saying

“that really ought to go in to a [subfolder].” When looking at the Desktop, he noted a document that was no longer needed, saying *“that shouldn’t be there, I’ll get rid of that”* as he deleted it.

Folder structure limits

Many participants remarked that reorganisation activities such as splitting a folder into multiple folders or creating subfolders were prompted by a folder reaching some limit.

Alex says: *“what normally happens is my Desktop fills up to a point where it’s unusable, and then I just pull out all the unnecessary stuff and pop it into a cleanup folder.”* Harriet also explains that *“when I get too much in a folder, I’ll create a subfolder in it, I mean, that’s where that old admin came from, I realised I had multiple so I shifted all the old stuff into a subfolder, so if I’m finding a folder is getting cumbersome, I’ll simply create a subfolder and shift everything in it, or clean it out.”*

Jack doesn’t have any rules for when to clean up a folder, but says:

“while I might not do that in a proactive way, in practice I’m reacting to the fact that it’s building up and I’ll think well I’ll subdivide at this point. And that could be something in the order of ... and again, it will depend upon the topics that are there. No point in differentiating them if there’s only one topic. If there are two quite distinct topics, you might think that. And that might be something on the order of 10 to 15 I guess.”

4.2.2.4 Windows XP Document locations

The two primary storage locations in Windows XP are My Documents and the Desktop. **Section 2.3.2** has more information about these locations.

My Documents as primary storage

Candice, Frank, Edward, Ina and Jack all use My Documents as their primary storage location. Edward mentions that he also has a working document folder for documents related his main research area, but everything else is saved into the My Documents folder. For these participants, this is the section of their documents that they spend the most effort organising, and which has the most extensive subfolder structure. Many of them, in showing this section of their files, would move files around as they did so, or talk about plans to reorganise or move things at a later point.

Gail mentioned that she uses My Documents as her primary storage location at home, but at work she uses the folder above it (her C:\Documents and Settings\<username> folder) because she was told that it was automatically backed up. The reason why that folder is backed up is so that all the subfolders of it can be backed up, including My Documents, the Desktop and other user settings folders. Thus Gail’s use of the folder above was actually based on a misunderstanding of the backup situation.

My Documents as “junk” folder

Damien describes the My Documents folder as his *“generic junk folder”*, saying *“it’s just where I put stuff that I’m too lazy to think of a category for, and then eventually it will annoy me and I’ll go in and you know, put it somewhere.”* His primary document storage location is his Desktop.

Although Ina uses her My Documents folder as her primary document storage, she also uses it as a junk folder, saying that she is *“quite ad hoc to begin with. I often just save it into the open My Documents and then I figure out where it all goes later. So if it’s something I’ve been working on recently then normally what I’ll do is check in the unfiled folder first.”*

My Documents as archive

Alex and Brett use My Document as their archive of documents that they wish to keep but no longer use or work on.

Brett explains his process: *“Well, pretty much anything that I’m using right now or have been using recently are the files on the Desktop. Periodically these files get dumped into folders on the Desktop. Periodically those folders get dumped into, I guess other folders either in My Documents or at times onto the C drive.”* He observes that most of the stuff in there is old material, with the most recent document being from the previous year. He says he sometimes goes back to this material, but that it is old files that were used in the past. He does note that there are a couple of specific folders inside My Documents that contain current items, mostly because it is the default save location for certain applications.

Alex’s use of My Documents is more complicated. He says:

- *“My Documents is for files that actually have a value and that I might want to revisit later on”*
- *“This one [opens another file], is also a half complete report I did for the ... survey with all the data in it. This sort of data I’m not inclined to delete but it also doesn’t qualify as one of My Documents.”*
- *“I try to keep all my working documents in My Documents.”*
- *“If it’s really recent, it’ll be on my Desktop, if it’s recent but I’ve finished with it, it’s probably in my My Documents.”*
- *“What ends up happening is that if it’s a working document that I come back to time and again or something that someone’s emailed me, it normally finds its way onto my Desktop, but if it’s something that I’ve created for people with the intention of working on it, like slowly, if that makes any sense, it goes into my My Documents.”*
- *“You see the Desktop and My Documents are pretty fluid for me.”*
- *“If you see here [points inside My Documents], I have created some subfolders to try and hive off some areas of my life so to speak, like Project X. I once created an Project X folder for the Project X things, but you’ll notice obviously that hopelessly, these Project X files are here, but there’s heaps of them on the Desktop, and in fact if I were to, oh gosh, this critical view of my file system. If I were to ... the most I would do would be to go to my Desktop, search Project X [does so], and I wouldn’t go so far as to copy them over directly, I would go open containing folder and go, well these ones, Project X reports, I can add those onto the [moves the files from Desktop into the Project X folder] but then, my fear would immediately be that now I’ve moved them off my Desktop, now I won’t know where they are. Even though it’s not much good on my Desktop either.”*

He has tried to make a distinction between the Desktop and My Documents with regard to the permanence of the files they contain, but without having a clear mental distinction between the two types of files, it hasn't really worked for him.

Desktop as primary storage

Damien and Edward keep their documents on the Desktop but rather than using the Desktop surface itself as a working area, these are organised into a permanent folder structure. Damien's system was most extreme, with his Desktop appearing to be almost empty, when it in fact contained his entire document collection organised into three top level folders. He explains, *"I can't stand clutter on my Desktop"*. He used to keep his documents in My Documents but found it more convenient to have them immediately accessible on the Desktop: *"it's a lot more efficient for me just to be clicking here"*.

Edward splits his documents, moving the more frequently used folders to the Desktop, and leaving less frequently accessed ones in My Documents: *"for the more useful ones, I put it here because it is more convenient for me to open again and again."*

Brett also has some of his documents primarily stored on the Desktop, while others are in My Documents or other locations on his hard drive. But he draws a distinction between the document collections that he stores in folders on the Desktop and those documents that he actually works on that are on the Desktop surface: *"But generally these folders are only used for things like the coursebooks which are generated by semester and the exams, which are generated by the end of semester as opposed to normal working documents."*

Desktop as worktop

Alex and Brett create most of their files directly on the Desktop and save all incoming documents to the Desktop. They allow files to accumulate there until it is full or nearly full, and then remove them all to another folder (often itself on the Desktop). Alex explains his process:

"These [points to some folders called cleanup1, cleanup2 etc] are basically past Desktops, I keep ... I don't like to throw things away, so even though all this information is old Desktop information, [opens a cleanup folder], even though all this information is essentially snapshots of old Desktops, I archive them in here so that I've always got access to them and I have had occasion to go back to them. What normally happens is my Desktop fills up to a point where it's unusable, and then I just pull out all the unnecessary stuff and pop it into a cleanup folder, so you'll notice if I say, go to cleanup1 [opens the folder], there's these Excel files are reports I did."

Any recent files will always be on Alex's Desktop, and he'll find it by scanning through the list. *"Normally I keep it in order of creation so that the current file is the last thing on the list so I don't have to look very far, and [I] travel up the list to see the older files."* Alex also refers to items being moved into cleanup folders by saying they *"fall off the scope"*.

Anything works on is placed on his Desktop: *“I’ve had this on my Desktop on and off for weeks. It goes off my Desktop when I believe it’s fixed, then [a colleague] sends it back to me in email and it goes on my Desktop again and I work on it. It goes back when I think it’s fixed.”*

Having documents on the Desktop provides some visibility that reassures Alex they won’t get lost. When talking about the possibility of moving files from the Desktop into My Documents, he says *“my fear would immediately be that now I’ve moved them off my Desktop, now I won’t know where they are. Even though it’s not much good on my Desktop either.”*

Brett has a fairly similar way of working: *“Well, pretty much anything that I’m using right now or have been using recently are the files on the Desktop. Periodically these files get dumped into folders on the Desktop.”* Some of the folders on the desktop are temporary folders similar to Alex’s cleanup folders, but others are more permanent folders embodying his primary storage.

Desktop for reminding

Candice and Damien both mentioned that having items on their Desktop reminded them about certain tasks. Candice had a document she’d received a while ago that she was intending to go through when she got a chance, so she placed it on her Desktop so that when she had time, she would remember to look at it.

Damien keeps a single file on his Desktop which is a journey log he needs to claim travel expenses. He explains it is only there because *“I always forget to record my ... trips because I used to do it on paper [points to his notice board], so if I see it, I’ll remember, because it’s quite important that we record things properly.”*

Desktop for reference

Candice keeps documents on the Desktop for frequent reference. These are frequently accessed and infrequently updated, although she does update them every now and then over long periods of time, for instance, a list of frequently called phone numbers.

Desktop for convenient save

Frank uses the Desktop as a convenient staging area for documents downloaded from email or the internet until he looks at them and then discards them. *“If it’s an attachment that I think that I probably won’t want to save long term, then I just save it to the Desktop and open it there.”* Alex also mentions that frequently his software development environment asks him asked to save a temporary file, and he saves it to the Desktop because it is convenient. He also notes some files downloaded from the web that were saved to the Desktop as a convenient location.

Edward also reports downloading material and putting it all into a folder on the desktop because it is a convenient save location. He later evaluates them and moves the most useful ones into a more permanent location, with the non-useful ones being discarded.

Desktop for shortcuts

Gail and Ina don't keep documents on the Desktop at all, instead only using the shortcuts that many programs automatically place on the Desktop. They seldom add anything to the Desktop themselves, as Ina says, *"unless it's been by mistake."* Gail doesn't actually want many of the shortcuts on her Desktop, and in fact, didn't even know what some of them were for. But because she didn't create them and therefore doesn't own them, she doesn't feel she has the right to delete them.

Damien deliberately removes all of the system-generated shortcuts from his Desktop. There are two that he cannot delete, and he places these in the bottom left where they will be least intrusive. He does keep a shortcut to one of his most frequently used locations, a shared network folder.

Jack doesn't put shortcuts on the Desktop, preferring to use the Quick Launch shortcut menu on the Windows XP Taskbar. Similarly, Harriet uses the Microsoft Office toolbar to contain all her shortcuts. She finds this a much more direct way of accessing her files than using the Desktop, which is usually obscured by the documents she is working on: *"why go out to get to the Desktop to go somewhere else. I can do it all without, I can cut stages out of the process by actually doing it directly from here, at least that's how I function. So, no, I find the Desktop horrendously frustrating. I much prefer going direct."*

No use of Desktop

Gail, Ina, Harriet and Jack do not keep documents on their Desktop. Gail's and Ina's Desktops contain shortcuts which they use, while Harriet's and Jack's contain shortcuts that have been placed there by applications, but which they do not use. Gail says she hates the Desktop and never uses it: *"I just detest the Desktop."* She finds there is nothing the Desktop offers her that she cannot more efficiently accomplish some other way.

Spatial use of Desktop

There were a number of references to spatial aspects of their documents during the interviews. This most commonly arose on the Desktop, with people reserving different areas of the Desktop for different purposes.

Candice uses the bottom of the screen for things she's currently working on or looking at frequently. The top left contains things she doesn't use very often, or hasn't used in a long time but would like to get to someday. And the rest of the Desktop doesn't have any special meaning.

Damien keeps two system generated icons that cannot be deleted in the lower right corner where they won't be noticed, *"because you tend to focus on the top left quadrant."* In the top left, he has his major top level document folders and a document he uses daily. He has removed all the usual Windows XP icons that normally occupy this area.

Edward has documents grouped into areas on his Desktop, but the areas themselves have no particular meaning. He just likes keeping related documents grouped together separated from other

groups of documents. Frank also says he tends to group things, although acknowledges that *“they get ungrouped over time.”* He says this doesn’t particularly bother him, and *“I don’t feel forced to go back and put things in order if the Desktop has got out of order.”*

Ina doesn’t really use the Desktop, but does make use of spatial cues in organising her physical documents. When asked where she keeps her documents, she pointed all around her office and said *“around me.”* She explained that her physical documents are her reference documents and her computer is for things that she’s working on or modifying. She frequently pointed to different areas of the room when talking about different types of documents, for instance referring to a journal article she’s no longer working on as being *“up there”*, pointing to a high shelf. The more current ones are immediately behind her in a filing cabinet. When asked about improvements to document management support in Windows, she said:

“I suppose there is a way to do it that I don’t even know how but that if I’m working anywhere, I’d like to be able to pull up, like I’d like to be able to turn around and look at these files. I have to kind of find my way through. Like there’s no icon that I know of that’s up top that says check your files. It’s not like there’s a filing cabinet handy, I have to kind of move down a few layers or move across a few layers to get to it.”

Seven of the participants consented to having screenshots taken of their Desktops. These screenshots are shown in **Figure 23** below.

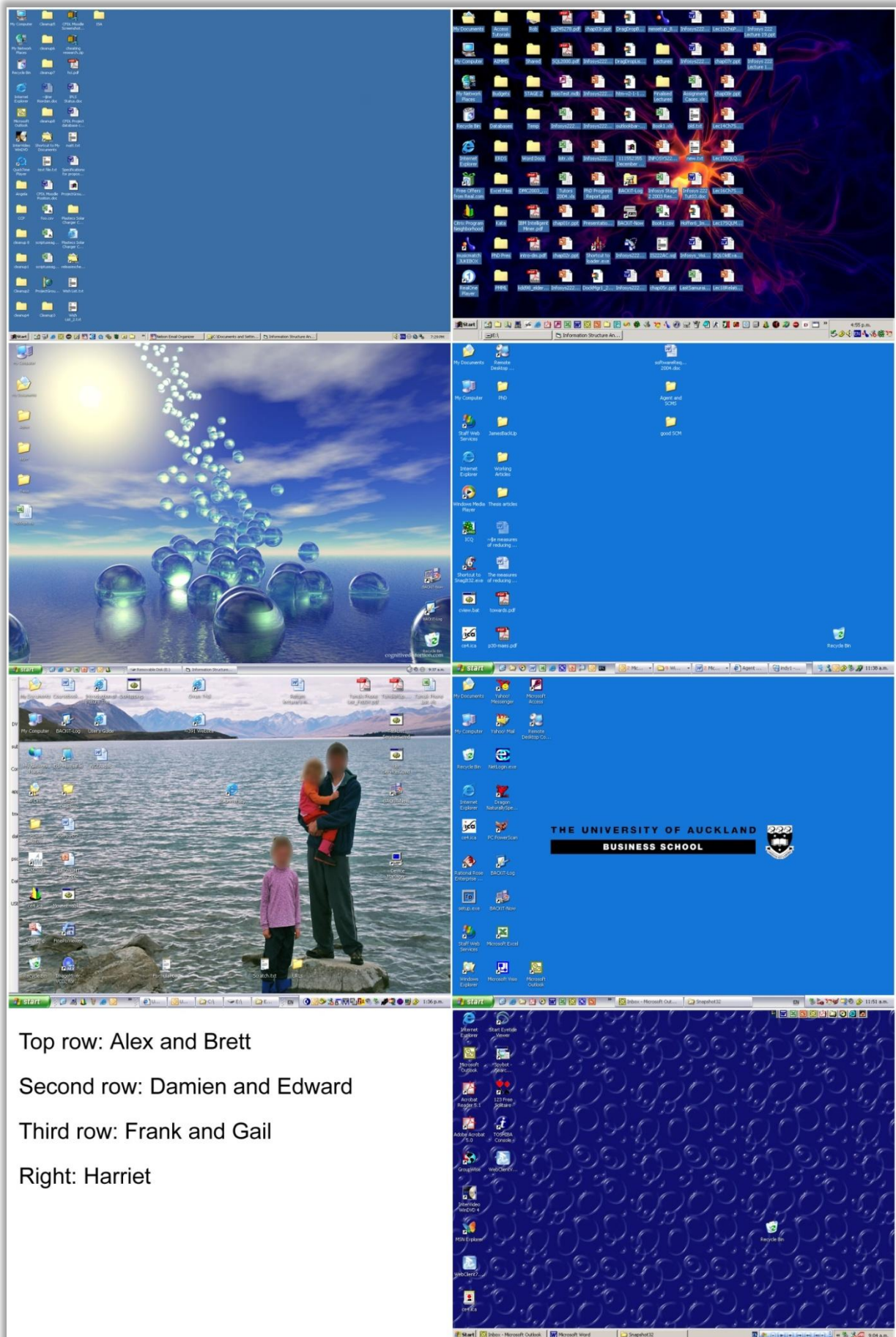
4.2.2.5 Finding Documents

This section describes all the concepts and behaviour arising from locating, retrieving and reusing documents, including the views they use to show their documents.

Browsing

The majority of the participants reported that if they need to locate a document, they would browse through it in their folder structures. Brett calls this a *“manual search”*, saying *“Normally there’s only going to be a couple of places it’s going to be. One of maybe three places.”* Candice also says she’d go and look in her normal structure, and depending on what kind of file she was looking for, she’d know where to go and look. Damien also says he usually knows the whereabouts of all his files. He says *“I usually know where I put stuff.”* He doesn’t see finding his documents as a big problem except that it takes a little time: *“You know where they are but you still have to click through to get them.”*

Edward says he can browse back to the folder if he remembers it: *“Normally, I just go to ... if I remember clearly where I put there, then I go to for example My Document folder to find it, or other folder to find it.”* Frank makes this process easier on himself by having a standardised organisation structure across all his computers and between his email and document folders. Because he is so familiar with his folders, he says *“if I’m looking for something, the zone of possibilities is quite small, and I can find material relatively efficiently.”*



Top row: Alex and Brett

Second row: Damien and Edward

Third row: Frank and Gail

Right: Harriet

Figure 23: Desktop screenshots of seven participants

Gail usually knows where her files are stored because she feels the structure she has created makes sense for her. Harriet finds it easy to find files again: *“Most of the time I actually know what my files are. It’s scary, but I can usually find what I want immediately because I take care in using consistent naming.”* She feels that because her folders and files are all named sensibly and obviously, even a stranger to her file system should be easily able to find things by browsing through her documents.

Ina takes pride in being able to locate her documents immediately, both physical and digital documents. She says files are in the first place she looks about nine times out of ten, and says that *“People say that they are surprised that they can come into my office and they can ask for an article and I will know where it is”*.

Jack browses for document if he can remember exactly where it is, otherwise he resorts to a search tool. Similarly, Alex will find a document by browsing if it is on his Desktop or if he knows for certain it is in his most recent cleanup folder, otherwise he will search for it.

Searching

Alex is a very frequent user of search. On numerous occasions during the interview when he wanted to show some particular aspect of his organising scheme or naming scheme, he would open the Windows XP search tool and search for the files or folders he wanted to show. Almost all his searches are based on locating keywords from the file name:

“I normally know, for example that, if [colleague] asks me what stuff isn’t done yet for Project X, I know that I had a document in somewhere here that was called ‘Project X to-do’, so I search for to-do and the document will immediately pop up. The trick really is in naming files uniquely enough that search works, so that’s why you’ll notice that my file names are rather heterogeneous. Apart from the fact that they all seem to have the term [department] in them, they’ve all got different names, it’s very unlikely that you’ll see two files with similar names, even in different directories.”

He mentions that he would sometimes like to search for the contents of a file, but because most of the documents he works with are Microsoft Office documents, they cannot be searched, and so he usually relies on searching the file names.³ While he never searches by date, he does often sort his results by date:

“The reason I don’t search by date is because dates for me are very relative, so I can’t give an absolute timeframe in which the file was created. I’ll know it was earlier than this file, if I sort by modified I can go, well I know that the file I wrote was before this and after this, so somewhere between here I could do a binary search, but I can’t go in absolute terms it was this month in which the file was created.”

³ This is not correct. It is possible to search the contents of Microsoft Office documents from the standard Windows XP search tool.

He very rarely searches by size, and would only do so if he knew he was looking for a particularly large file. He doesn't search for file type using the search advanced options, but by putting the file extension in as one of the keywords in his file name search.

While Frank prefers browsing if he can remember the location of a document, he very often resorts to searching for it: *"I told you that sometimes a mess, I try to find some file, so I use a lot of search."* His preferred method of searching is to first try keywords from the filename, and if that doesn't work, to move on to searching text within the file: *"I try to remember the keywords. For example, first of all I try to search the documents, the folders, by remember some keywords for the folder name. If unfortunately I could not find that folder or the file name, I use search for the text within the file, to search that. Sometimes it work very well."*

Jack uses Copernic Desktop, a third party search tool in order to locate his files. While he does browse to something if he knows exactly where it is, he frequently turns to Copernic to find his documents. He describes it as a *"Godsend"*, because he doesn't have to adhere to his clear directory structures but can just *"shove files anywhere"* and know that he can *"just pull them back by searching for them without having to know where they are."* He primarily searches by file content, not the file name: *"It's that content which I tend to remember, people maybe with more logical minds or more structured minds will recognise that it's this type of topic and therefore it's in this folder for this topic, but because my folders got rather messed up, and I can't be bothered to get back and sort them out, I now just don't worry too much."*

He also appreciates the preview feature that Copernic offers (which the standard Windows XP search doesn't have), allowing him to easily see if he has found the right file or not.

Brett rarely searches for documents, doing so only if he can't remember where he put a file, often because he's moved it somewhere else on the computer. His searches usually don't narrow down the location, because *"if I'm using search in the first place it's because I don't know where something is, so I'll search the entire hard drive."* He normally searches using keywords from the file name, and if he knows the type of the file, will include the extension as a keyword. Candice and Gail also say for them search is a last resort if they haven't been able to find something by browsing documents, and they'll also use keywords from the file name.

Frank also rarely searches his documents: *"Occasionally I do, but by having a standardised structure, more often than not the document is where I think it should be."* His use of replication tools also means when he does search, he only needs to search one document collection, rather than wonder which of his multiple drives might contain the document.

Damien almost never has to search for something in his own documents: *“I don’t use the search function as such, because I usually know whereabouts they are.” “But I suppose, rarely, I’d say I use the search, but sometimes, I suppose you might lose something.”* If he had to search, it would be using keywords from the file name.

Harriet has a similar perspective, saying the only time she needs to search is occasionally when *“I’ve saved an attachment without paying attention, yeah, occasionally I’ll do that. It saves to the default temp folder and I think, ‘I’m sure I saved this in the right place’ and so I’ll go and search and it’s invariably in the temp download folder.”*

Ina has never needed to search for anything:

“My husband tells me I should use it but I haven’t lost anything badly enough to need to do that. I’m pretty good at finding what I’ve lost. Really I could count on maybe one hand if I’ve ever thought ‘where the hell is that file?’. I just haven’t run into that. And that’s like forever, so I don’t think I know how to use that Search File properly. Whereas the way my husband files, he’s always searching.”

Failure to find

Alex reports frequently being unable to locate his documents by browsing: *“I’m constantly going, oh, I had a file on my desktop the other day that had such and such on it and it’s not there anymore and I’ve squirreled it away into one of my cleanup folders and my cleanup folders mean nothing, cleanup1, cleanup2, cleanup3, cleanup4 means nothing so I have to do a search through my desktop to find the file.”* He says that he will sometimes generate or download a file more than once because he can’t find the original version, but that usually he manages to find what he’s looking for. He says the trick is in unique file naming to make his search work better, and also notes the only reason he bothers to reorganise anything is to make search work better. If he’s unable to find the document by searching, he’ll look in his email and see if he can find the document as an attachment to a message he sent or received.

He says it is infrequent that he completely fails to find a file, but that sometimes *“the cost may outweigh the benefit.”* During the interview, he several times searched for a file to demonstrate a file that had version numbering through multiple versions, and had to formulate three searches before he found it, saying that was the *“story of my life.”*

Edward reports occasionally failing to find a document he was sure he’d written. Frank also reports being unable to find documents because he’d been searching on the wrong words or phrases. Thinking of other keywords to search on normally eventually yielded the file, but *“I think there was one occasion where I realised that I’d actually dumped the file so it was never going to find it, but that’s a sort of eureka, now I know what’s gone wrong sort of thing.”*

Sorting

Many participants referred to the use of sorting to help them find documents. In some cases, this was sorting the results of a search in order to find the file they were after. In other cases, this was as part of a browse strategy, with sorting being a way of grouping documents within a folder in order to be able to find particular subsets of files easily.

The default details view in Windows XP provides four columns (Name, Size, Type and Date Modified), with an additional 34 attributes available for display. Most participants reported that while they had a single sort preference, but changed between sort attributes depending on what they were trying to locate. For instance, Alex describes how he changes the sort attribute depending on what he knows about the document he is looking for:

"If I'm looking for a particular file here, sometimes it's useful to sort by type, cause I know it's a text file and I can go to text files and then find it immediately, [foo].txt for example. Sometimes I know that it's the most recent file so I scroll down to the bottom and there it is. Sometimes it's by name and sometimes it's the biggest file that I'm looking for, so I can know relative what's the most efficient way for me to find it."

Damien, Brett, Candice and Gail said their documents were normally sorted by Name. Damien never sorts by any other attribute, Brett would sometimes sort by date, and Candice would sometimes switch to either date or file type. Damien did mention that at home he occasionally sorts by file type to group different file formats together, but seldom has a need for this at work. Brett has no need to sort a given folder for file type because he usually ensures that different file types are placed in separate folders.

Edward reported equally sorting by name, by date and by file type. He mentioned sometimes naming files in order to force a certain sort order, for instance:

"For example I found that at home I name my backup for my [project] by the date. It's easier to get it back, and also the date is not the date we use in European countries, use the date sequence like year month and the date and the time, and then I can sort it out and find out the latest version easiest."

Alex also does the same thing, formatting the date a particular way to facilitate sorting. Harriet noted that she also names her files with sorting in mind: *"That's one thing I am careful with though, because it's a 12 week course, I always put the zero in [Module 01 not Module 1] so they actually stay in order."*

Jack noted that *"the important things to me are the name and the date modified."* He says that the date view works well for him because *"you're often working on a particular topic, or particular project and it's a recent one and most of the files that you want will be of recent date and so that's good."*

While Alex changes his views when looking for a file, the rest of the time he uses a predominantly date based sort. Items on his desktop are arranged in the order they were added: *"normally I keep it in*

order of creation so that the current file is the last thing on the list so I don't have to look very far, and they kind of like travel up the list to see the older files." Alex prefers date based sorting rather than date based searching because he prefers to work with relative rather than absolute dates.

Alex was the only one who mentioned regularly sorting by size. Frank does use a third party program to find out the sizes of his folders, but he explained that it was mostly to find large files taking up space on other computers he manages. Candice said that size was not a relevant concern for her and that she didn't think in terms of size and had no clue what size things were supposed to be.

File views

The two views most commonly used by participants were the list and details views, with icons only used by one person, and disliked by three people. **Section 2.3.2** shows the differences between these available views in Windows XP.

Edward's preference for the details view is *"because sometimes I need to check the size of the file and also the date. And also sometimes I try to find out or organise my file by using the type, so these three attributes are quite often used by me."* Brett also says he uses the detailed view whenever he can, and Candice says that if she ever encountered a folder in another view, she'd change it to details. Gail's preference is due to the ability to see more information at a glance.

Jack's preference was for details view:

"I much prefer to see a details list of files, I don't like icons." He didn't feel other people necessarily shared this preference however, saying "most people do like icons and if they were to say because of that they were going to get rid of the details list I would think that was a retrograde step." Alex also explains that because he so often works with lists, he always uses the details view: "I hate viewing it in terms of icons."

Harriet also says *"I hate icons,"* preferring normally to use the list view to see her documents, although *"occasionally I'll look for details, particularly with my online course, I'll sometimes look at that with details because that has a lot of things I had scanned into PDF form, and in details I can pick up its size."*

Damien had the details view in place for all his documents, but explained that he normally uses the list view. When asked why it was in details view, he replied:

"Because I haven't bothered changing it. Cause I don't have a ... OK, well, at home it's very different, cause at home I have a lot more information. That's actually, yeah. I think cause I don't have enough in here that it goes onto multiple screens. As soon as it hits multiple screens then I go to list so I can see everything. But here I probably have it on detail just cause it hasn't annoyed me. Hmm that's actually interesting, because I don't actually use list the most here, I use details, quite right."

Ina was the only one to use the icons view at all, and uses a combination of list and icons depending on how she is accessing her files. If she is using My Computer, everything is in icons view. However, her

primary mode of accessing files is through the Open/Save dialogs in Microsoft Office applications, and there, due to the small size of the window, she changes it to list view so she can scroll through more easily. Her use of My Computer and the icons view is usually reserved for when she is doing a reorganisation.

Tree view

Seven participants used the tree view when exploring their documents while three did not.

Of the three that did not use the tree, both Alex and Harriet still had the tree visible in the left hand pane (the default setting in Windows XP), but did not use it to navigate, instead double-clicking through files in the details or list view in the right hand side. Ina did not have the tree visible at all, and mostly uses the Open/Save dialogs in Word to navigate through her documents. Since Ina tends to have a very flat structure with few subdirectories, she normally only needs to click once on the folder in order to see the file. Alex says that he tends to have very flat structures and doesn't use the tree. He describes himself as *"tree averse"*, pointing to his My Documents folder which contains 32 folders and 170 files and saying *"that's ridiculous, how can any sane person possibly cope with that? That much vertical stuff."*

Brett, Candice, Damien, Edward and Frank all used the tree to navigate through their documents during the interview, but didn't comment specifically about it. Jack used the tree and commented that he frequently navigates down the tree to find stuff.

Gail commented that she was a bit annoyed at having to click down through the several levels of the hierarchy in order to get to the folder where she stores her documents. She starts from My Computer, and then expands C Drive, then Documents and Settings, then her user folder and then the folder she wants to work with. Because she doesn't use My Documents, the shortcut automatically provided by Windows XP doesn't help her access her documents faster. At the end of the interview, the researcher offered to show her how to make a shortcut directly to the folder where she kept her documents, but she declined, saying she'd rather stick with the way she knows.

Necessary metadata for successful finding

Alex mentioned needing to know the keywords from the likely name of the document, the type of file he was looking for and how recently he used the document. He would normally use keywords from the name and file extension in his search criteria, and then sort by date in order to narrow down the search to the document he was looking for.

Similarly, both Edward and Gail indicated that they need to know some keywords from the name of the document, including the file extension. They both don't mention date, but if they fail to find it by searching on name, they'll try searching on some possible file content instead.

Ina and Harriet both mentioned needing to know the type of file they were looking for first, because their primary method of finding documents involves starting from within the appropriate application and using the Open File dialog to browse for the document. Hence they need some idea of the topic the document pertains to in order to choose the correct folder, and they an idea of what they might have called it in order to recognise the name when they see it.

Brett browses through Windows Explorer, but he does narrow his search by file types by the fact that he has separate folders for different types of file. So as long as he knows the file type and the approximate topic he might have filed a document under or the course it was related to, he can browse to the file.

Candice indicated that before she could find a document, she would need to know the genre of document she was looking for, explaining that if it were a final exam she'd look in one place, whereas an outline or documentation would be in different places.

Frank needs to know the year, semester and course that the he was looking for relates to, and then the genre of document, such as whether it was an outline or a coursebook.

Jack's searching technique operates exclusively on keywords and phrases from the content of what he's looking for, with a possibility of sorting by date if he has to choose from multiple possibilities.

Recent vs. Old documents

Both Candice and Ina indicated that their find tactics would be different for recent files compared to older files. They both use the Recent Documents menu in the File menu of Microsoft Office applications to open one of their four most recently used documents. For any document that they knew was older than that, or that wasn't there, they would use their normal browsing process.

If Alex thought the document he was retrieving was a recent one, he would scan his Desktop for it. If he was sure it was quite recent, he might also scan his latest cleanup folder for it, before resorting to search if he couldn't find it. If he knew it was an old document, he would just go straight to search.

Frank and Gail both said they would use their standard browsing technique regardless of the age of the document, although Gail noted that she might be less likely to find an old file and therefore more likely to resort to search.

Harriet indicated that her tactic would be the same regardless of the age of the document, although she noted that *"when I open a document, I usually try and finish what I'm doing with it, so it would be unusual that I would be working on a document that I've used within the last 4 or 5 times because I obviously didn't do the job properly the first time."*

4.2.2.6 Retention

Retention encompasses the decisions about whether to retain or delete a particular document or folder. It includes decisions about whether, when and how to back up documents for safekeeping, as well as examining reuse of retained material from their archives.

Deleting

While all participants have sometimes deleted files, there was a very clear division between those who were inclined to delete and those who were not.

Deleting decisions

Alex says of his documents: *"I'm a hoarder; I'm loath to delete them"*. He repeats this sentiment on eight different occasions during the interview, for instance, *"I just don't like to delete anything"* and *"I don't delete them because that's the way I am"*. Edward also is very reluctant to delete, saying *"most of the files I don't delete them, I just save them to archive"* and *"I still keep it, I duplicate the thing, just in case."* Gail also reports keeping all her files, only occasionally deleting temporary files or old versions.

Ina tends to keep everything except for old versions of documents that relate to projects that have ended. She keeps the final version, but regards the older versions as temporary files and not something that needs to be retained any longer. Harriet deletes everything more than two years old, but she has a thorough system of monthly backups on CDs which are kept at multiple locations to ensure that she never loses anything. She says, *"I don't often delete documents, unless it's a global delete because I've archived it to some annual disk or something."*

In contrast, Damien prefers to purge documents from his file system: *"I'm a deleter, I don't like archiving things"*. He either immediately permanently deletes the document, or immediately empties the Recycle Bin. *"I don't like keeping old documents,"* he explains. Candice similarly deletes any document she thinks will have no future use, and immediately empties the Recycle Bin when she deletes documents. The exceptions for both are documents they are required to keep for audit purposes.

Brett and Jack also report deleting things if they can't foresee any future use for them, but keeping them in the Recycle Bin, *"just as a precaution"*. Frank also deletes files that he no longer regularly uses, but he does always make sure there is a backup copy first.

Use of Recycle Bin

Alex doesn't use the Recycle Bin. Since he is very reluctant to delete, and has an archive process, when he deletes a file he is absolutely sure he really doesn't need it, and permanently deletes it. Candice generally does a two step delete. She deletes the file sending it to the Recycle Bin and then goes into the Recycle Bin and permanently deletes it from there. She says *"I just want to get rid of it, because I regard it as being no longer relevant."* Damien also empties the Recycle Bin regularly, *"daily"*. He emptied the Recycle Bin after deleting a couple of documents during the interview.

Damien has sometimes been a little overzealous with his deleting:

“yeah, sometimes I’ve regretted deleting things, but normally, I think that was more when I first started work, I realised, oh, I might need that. Now I’ve sort of got more experience I know what I’ll need again. So if I think I might need it again, I keep it. But again, it’s still, sometimes you delete stuff you don’t want to delete, but hey, that’s life.”

Brett uses the Recycle Bin in case he ever changes his mind about deleting a file: *“I do use the Recycle Bin, but I don’t really empty it, so even if stuff’s there I know that I can still get it back.”* He has had to retrieve stuff from it on occasion, but says it is very rare. Jack also uses the Recycle Bin this way, and says that retrieving a deleted item from it is very rare because *“I’m careful enough when I delete to only delete those things that I know are very unlikely to be needed again.”*

Gail and Ina also use the Recycle Bin. Ina rarely needs to retrieve from it, *“maybe a couple times max.”* She does clear it out every year or so, and says it doesn’t usually occur to her to look in there: *“I’m the kind of person that if it’s not there, I don’t even look at it.”*

Archives

Many of the participants spoke about having archives of material that is no longer active. Amongst the academic and academic support staff, this frequently consists of material related to courses taught in previous years and semesters.

Alex reported in the past having archives of documents in zip files but says, *“I don’t tend to do that now because I’ve got such a massive hard drive that there’s no need. I just let the data atrophy.”* He now has locations on his computer where he puts documents he no longer actively uses:

“These [points to some folders called cleanup1, cleanup2 etc] are basically past desktops, I keep ... I don’t like to throw things away, so even though all this information is old desktop information, [opens a old desktop folder], even though all this information is essentially snapshots of old desktops, I archive them in here so that I’ve always got access to them and I have had occasion to go back to them.”

Alex reports needing to go into these archives every couple of weeks or so.

Brett has dated archives that are snapshots of his system at various events, such as getting a new computer, or a software upgrade. However, he says he only needs to dip into these a couple of times a year. Damien has a set of archive folders for each year into which he moves the previous year’s documents at the end of each year. He revisits these usually at the start of each semester, to get the previous versions of several recurring files to use as a template for the current semester. He says the need is *“normally when you’re setting up for the beginning of the semester you look at what you did in the past, but other than that it’s very rare.”*

Harriet also has a system of archive folders for each year, and also usually access her documents every semester or so, to pull out resources that might be useful for the current semester’s teaching.

Frank frequently accesses his archives: *"It's surprising actually a document from 3 years back suddenly you want to refer to it again."*

Backups

Frank, Gail and Jack all mentioned relying on the University's IT department taking regular backups of their systems for them. In fact, Gail places all her documents in the folder above My Documents because she was told that this is the folder that gets backed up and wanted to ensure she took advantage of the backup scheme.

Edward, Frank and Harriet take their own backups using a manual process. Edward backs up to a variety of different media, putting copies on CDs, DVDs, external hard drives and network drives that he has access to, as well as often having a copy of important documents on a USB drive. He usually dates the backups so he knows when they were taken. Frank uses CDs for his backups, usually having a backup for each year.

Harriet keeps her documents in a specific folder (C:\Harriet's Data) specifically to make it easier for her to take backups. Whenever she wants to backup, usually every week or so, she just drags the entire folder to her CD burning software to make a backup copy on rewritable CD. She keeps the most recent backup in her office, and the previous backup in her handbag, swapping them over every time she overwrites the older backup. Every now and then she'll also make a separate backup to take home: *"that's sort of my catastrophe backup."* She writes dates on the CDs so she always knows which one is the more recent one.

Brett, Frank and Harriet all delete folders and files once they are satisfied that they have been backed up. Edward doesn't, keeping a copy *"just in case"*. As he explains, *"because basically I understand that CD is not reliable either."* Ina also makes sure she keeps everything around in order to be able to access it faster, taking pride in her ability to quickly find things that other people ask her for.

4.2.2.7 Problems

Versions

The issue of document versions has not come up in the literature before now, but was raised by the first few participants as an issue they found in their document management practices. It was specifically asked about during subsequent interviews.

During the tour of his file system, Alex observed that he had multiple versions of a document in various folders and also in his email. The versions were numbered with a v1, v2, v3 prefix. He initially said that he thought version v4 was the most recent, and used a file system search to confirm that. However later he opened his email client and noticed that there was a v6 in there which was actually the latest version of the document. He noted that he himself wouldn't use a version number scheme, but that others began it and he *"just followed on the convention they used."* He personally says he uses

the Track Changes feature built into Microsoft Word so that he doesn't end up with multiple versions in different files.

Candice and Ina predominantly use dates to track versions of a document. Both also sometimes include the name of the most recent person to change the document in the filename. Candice also tries to keep the older versions out of the way by copying them in to a subfolder called 'old' so that only the most recent version was visible in its folder, whereas Ina will delete the old versions as soon as the project finishes.

Damien routinely includes a version number in his standard document naming scheme. *"This is something I've started doing recently, because [I] had a few problems with multiple documents, like when you've sent things to the lecturer and they've come back so it's a pain for me."* He also mentions that he tries to delete the older version so only the new one is around.

Jack also reported using version number, but not as a matter of course. He says *"... I do use version numbers if we're going through a rapid sequence of changes ... I do use versions but not frequently"*.

Frank and Harriet do not keep multiple versions of a file that is updated. Frank says *"it's something I've toyed with, but life is complicated enough"*. Harriet periodically backs up all her files and prints important documents, so she maintains a version history that way while only keeping one file.

Copies

Frank explains the problem:

"There's always a problem when you have multiple copies of a file, and you inadvertently do an edit on what in fact is an earlier version, and somewhere else on your many disk drives there's a later version that is really the one you should have done the latest edit on. It's not a good idea to have multiple copies of a file."

Frank goes to quite some lengths to avoid this problem. He keeps a copy of his documents on his work laptop and his home desktop, as well as a copy of some of the structure in a network drive, and some of the structure on a USB drive. This is partly for reasons of redundancy in case of loss of one set of documents, but also so he can always work no matter where he is. To keep everything synchronised he uses a freeware utility downloaded from the internet. The utility automatically makes sure all locations have the latest version of each document, and prompt the user for intervention if there is a conflict (i.e. if both copies have been independently edited).

Edward also runs into problems with copies in multiple locations:

"Just in case that I lose some files in somewhere, most of time I keep three copies, or even four copies. For example I have a memory sticker [USB drive], I keep some file there and also here [Desktop], and also my laptop and my home computer. Sometimes it's a problem because there are too many different kind of versions, so different kind of copies."

Edward tries to prevent the problem by trying to keep them updated: “I just update all the files, keep them synchronised immediately. If after a while I may forget it.”

Despite these precautions, he still sometimes discovered unexpected multiple copies during a reorganisation. He resolved the problem by checking the modification date to find the most recent, but kept both copies “*just in case*”.

Alex also encountered problems with unexpected multiple copies, usually discovered while searching for the document. He took no steps to resolve or prevent it, and would use the modified date to determine which to use.

Brett had a closely related problem in that he would deliberately duplicate documents in multiple locations, and then “*there could be some confusion as to where the last, most recent version is*”. His resolution was also usually to compare the dates to find the most recent one.

The other participants did not report this issue as being a problem for them, although Damien expressed a desire for a synchronising tool to ensure his USB Drive and hard drive remained synchronised.

4.2.2.8 Windows Features

Most participants were asked to identify the best and worst features of Windows XP related to document management, as well as identify any missing functionality, or anything they would like to see added.

Best feature

Alex said the ability to search for documents was the most useful for him. Brett said “*I only use the very, the basic, the basic structures,*” but commented that it was fairly easy to drag and drop things around.

Candice and Gail said that being able to create folders to group and organise things in folders was the best feature. Damien also said that being able to organise things in folders was essential, as well as the ability to sort and view details when required. Ina also pointed to the ability to sort things, and to switch from icons to details view when she wants more information. Jack also mentioned the ability to see the details of each file. The others couldn’t single out any particular best feature.

Worst feature

Damien found that his biggest problem was that files couldn’t easily be placed in multiple categories: “*you sort of put it in the one that it fits the most. You can, I know you can like put in shortcuts and stuff, but who’s got the time to do that, you know, to copy things from one place.*” He also mentioned the time taken to navigate the tree: “*it can take a while to get down through the levels, I mean, you’re talking seconds here but they add up. And it would be nice to think, zip, and oh, up it would pop, but you know, I*

mean, it's just a limitation of the way things are and you have click through the hierarchy to get to somewhere, you know." Gail also singled out the time taken to navigate down through the hierarchy as her main annoyance.

Frank noted that the inability to easily replicate structures was a problem for him – so much so that he devised his own solution from batch and script files. Brett identified having multiple separate file structures and locations as being a problem for him. Edward says that the fact that files can get duplicated in different places is the main drawback.

Alex mentioned the fact that documents in his file system have lost their context, saying *"I don't know who sent it to me, I don't know where it came from, I know that I didn't create it, I'm pretty sure that someone else sent it to me, and I saved it on my Desktop so I could deal with it. I don't even know if I've dealt with it."*

The other participants couldn't think of any particularly bad, annoying or time-wasting features of Windows XP.

Missing features

Frank suggested it would be nice to have dynamic folders like those Outlook 2003 provides for emails. Ina wanted a spatial dimension to her files, so she could in some way access files behind her and beside her to match the physical locations she has for paper documents. Jack couldn't think of anything he'd want to be added to Windows, since he was happy with his third party search tool and didn't think Microsoft would be able to do it as well.

The other participants didn't identify anything they would like to see added to Windows XP.

4.2.3 Summary of Interview Analysis

Participant's statements about their documents can be grouped into six main areas: Attitudes, Creating Files, Folder Structures, Finding, Retention and use of the Desktop and My Documents folders in Windows XP.

Attitudes

Under attitudes, participants generally indicated that being "well organised" is a desirable state. Those who felt they had attained that state expressed pride, and those who felt they hadn't expressed some guilt as well as the desire to reach it. Several spoke disparagingly of the perceived messiness of their folder structures. Some compared themselves with other people, or expressed a desire to know how their behaviour compared to others. Several mentioned that while they started with good intentions to be organised, over time they were unable to maintain it and slid progressively towards "messy". Comments about the hierarchical nature of file management in Windows XP were positive, with participants indicating they found the hierarchy intuitive and a good match for their mental

models. Some expressed variations on the theme of folders allowing them to put documents out of sight and out of mind, taking up no cognitive space. Others feared if that happened, it would result in loss of knowledge about the document, and hence loss of the ability to retrieve it. Issues of file ownership were important for two participants, and seriously impacted the document management practices of one participant. Some participants indicated they were willing to change document management practices if it would allow them to be more organised, while others were happy to stick to familiar and reliable practices, even if these were sub-optimal.

Creating Files

Most participants created files using the application that edits the file. It was rare for documents to be created directly inside a document collection, except in the case of reuse. Reuse of existing documents was extensive, with participants keen to make use of previous work. Document naming tends to be ad hoc and personal, with people trying to choose names they found meaningful to them. Naming conventions were never rigorous, but tended to be used for documents related to teaching, often with the expectation that the document would eventually come to live outside the folder hierarchy which provides metadata and context. Most participants reported creating multiple files for successive versions of a single document, using numbers, dates or people's names to distinguish different versions. Several participants inadvertently ended up with multiple copies of documents (which may be exact copies or different versions) and having to figure out which was the canonical file.

Folder Structures

Folder structures are sometimes created in anticipation of later being filled with content, just-in-time in response to the need to find a location for a document being saved, or to 'clean up' and group existing folders and documents together. People reported folders being named in various ways, with document genre, topic, time and course or project being common labels used. Several participants identified their own folder structures as fairly flat, disliking many levels of subfolders. Others mentioned synchronisation as being necessary to integrate multiple collections and locations. Reorganisations were common, with some people performing them periodically (often on an annual or per-semester basis) and others continuously. Yet others thought they should reorganise and wished they could but lacked either the time or the motivation to do so. Misplaced documents were common, with many participants finding documents and whole folders that should be moved or deleted in their system. Some spoke of having limits which prompted them to reorganise – either when the number of files in a folder reached a certain level, or when a certain level of 'messiness' was reached.

Finding

Two main methods employed of locating documents are employed: browsing and searching. Both rely on a combination of recall and recognition in order to locate the particular file. Browsing is a process of scanning files and folders and recognising the file being sought. Sometimes people using a

browsing strategy know exactly where a file is in their hierarchy and simply click through until they find it. In this case, they are relying more on their recall than recognition. In other situations, they are less sure of the file and its location, and open the most likely named folders and files until they locate it again. Because they themselves were the person who named and placed the file, they are likely to recognise it when they see it.

Searching involves using a search tool and specifying some search parameters and then selecting the result from the list. There is a Search feature built in to Windows XP, as well as other third party search tools available such as Google Desktop and Copernic Desktop. In order to search, the user must recall some information about the file being sought. Typically, search tools allow the user to provide keywords from the file name or content, the date the file was created or last used and the size of the file. The search can usually be narrowed down to specific folders if the user knows approximately where the file might be. The user is presented with a list of search results, which normally show the file name, file type, modified date, size and location. From this, the user must recognise which file it is they were seeking. The more the user recalls about the file, the less they need to rely on recognising it amongst the other search results. The less they recall, the more they need recognition.

Regardless of whether the person is using search or browse, they must view their documents, either within their folders, or within the search results. How they view these documents impacts on how easy it is for them to recognise the document they are looking for. In particular, the order in which the documents appear is an important step in the finding process, with the order often being changed to group files of similar type together or find the most recent files. Some participants used the tree view in order to browse down the tree to find documents, while others clicked through the folders.

Because of the trade-off between recall and recognition in the finding process, the age of documents is an important factor in the search strategy selected. People can generally recall less about old documents than they can about recent documents and therefore sometimes will use a different strategy to find them, generally being less likely to go directly to the document, and more likely to use a search tool.

Retention

Retention activities include decisions about whether and when to back up, delete or move documents to archives. Some people avoid deleting when they can, others prefer to remove items they no longer need in order to have a higher concentration of active and useful documents. Some people create specific archives, or folders containing documents that are no longer used, while others will leave them in their primary documents storage location.

Use of system-provided locations

People use the system-provided My Documents folder in various ways. Some don't use it at all; some use it as their primary storage, some as an archive and others as a temporary storage location. There are multiple ways in which people use the Desktop, from a working area, to storage, to shortcut area, to no use at all. Some of the participants use the Desktop in multiple ways simultaneously, or multiple ways at different times. For those who use the Desktop, the way they arrange things can differ, with some people taking advantage of the ability to create a spatial arrangement of items, and others not using this.

Participants mentioned the ability to group things into folders, and the ability to change views and sort items as being amongst the most useful features of the Windows XP's document management capabilities. The time taken to navigate using the tree was mentioned as one possible drawback, as was inability to replicate structures, and the lack of context provided by the file management tool. Most couldn't think of anything missing, with one request for more spatiality and one request for dynamic folders.

4.3 FILE SYSTEM SNAPSHOT

The goal of the File System Snapshot software was to capture data about the structure of the participant's folders and documents. Because this instrument was intended to be used with both the interviews and the survey, it needed to collect the data in a form amenable to analysing with a large number of participants.

At the time the research was conducted, no similar study had been carried out, and no software existed that could take a file system snapshot for later research. Software to take this snapshot was written for this research. The information it is designed to capture is shown in **Table 4** below. More information about the design and implementation of this tool is presented in **Appendix F**.

Table 4: Data captured by the File System Snapshot program

Element	Data Captured
Folders	Name Structure (position in hierarchy) Date Created Date Last Modified Date Last Accessed
Files	Name Path Date Created Date Last Modified Date Last Accessed No file content is captured

Most of the file system of a typical windows computer is not actually under the direct control of the computers' owner or user. The folder C:\Windows\ and its subfolders contains the files needed to run the operating system, and the folder C:\Program Files\ contains files needed to run all the programs installed on the computer. The user's own documents are buried deep in the file system in the folder C:\Documents and Settings\\. In addition, some programs also store their own data and configuration files inside the user's own Documents and Settings folder.

For the purposes of this research it was important to try and only capture the folders and files that have been created and are managed by the participant themselves. In order to accommodate this, the interface for the program includes the ability to select which folders to be included in the snapshot. By default the Desktop and My Documents folders were included, but the user is able to remove those, and also to include any other locations where they store their documents.

The participants were instructed to include only their documents, not operating system or program files. The snapshot was taken on their primary work computer, and could include network locations, but not other desktop or home computers. During the interview, the snapshot was run on their primary

work computer, and participants were asked to identify and include the top level folders containing their documents

The snapshot software took between 15 seconds and five minutes to run, depending on the size of the document system in question.

Figure 24 shows the user interface the file system snapshot software:

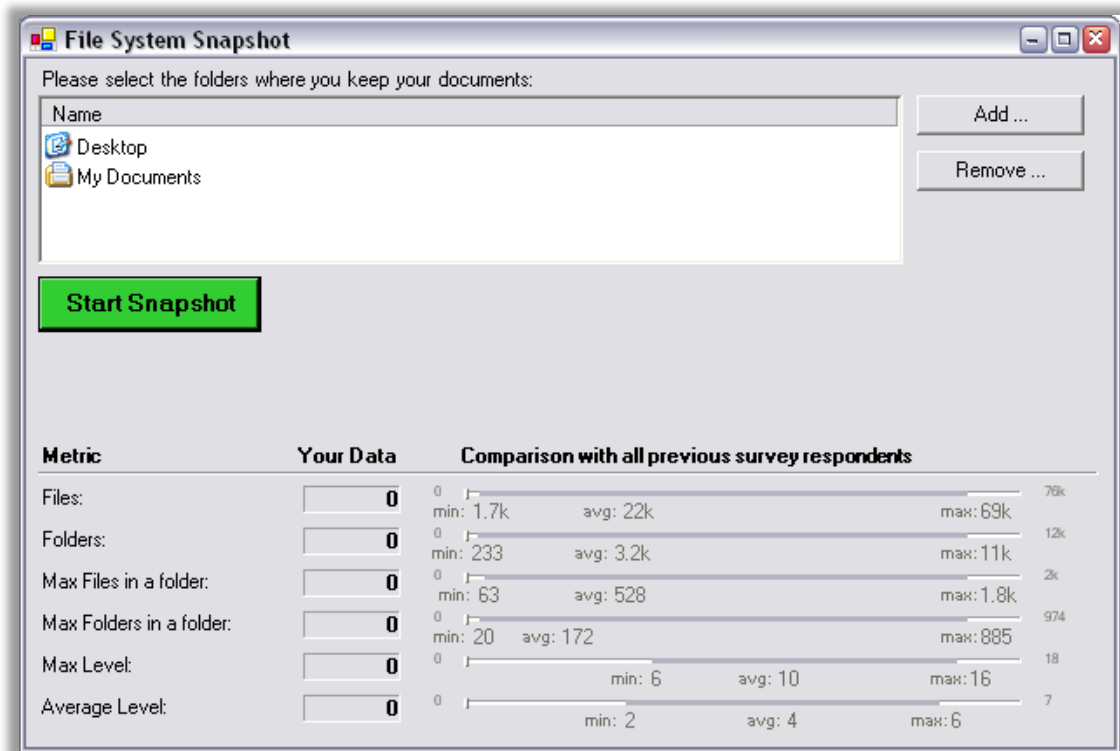


Figure 24: Interface of File System Snapshot Software

The metrics section at the bottom of the screen shows a comparison of how the current file system compares with previous participants on a number of major statistics. This was included for two reasons. The first is that it retains the user's engagement while the snapshot is being taken, since the snapshot can take several minutes depending on the size of the document collection being scanned. The second reason was to pique the interest of potential participants, since people are naturally curious about how their folders compare to other people. It was hoped that the display would both make it more likely that people would run the snapshot, and make it less likely for them to cancel it before completion.

The data captured by the file system snapshot was sent to a XML Web Service which stored it into a Microsoft SQL Server database. There was a backup system in place in case, for some reason, the snapshot software wasn't able to connect to the Web Service. If this occurred, it would instead save the data into a text file on the local machine from where it could be copied onto a removable drive and later imported into the database for analysis.

4.3.1 Metrics

From the data collected by the file system snapshot, a number of different metrics can be derived to describe the file system. These include the following categories:

Size. The size of the problem has an impact on the nature of the solution, since software that adequately supports the task of managing a few hundred files might not be sufficient for managing thousands or tens of thousands. The overall size of the system is measured by the number of folders and number of files. Additional measures include the number of non-empty files and folders, and the number of unique files and folders.

Depth. Systems vary in the level of nesting of folders. Shallow structures may only have one or two levels of nesting, whereas deep structures can have many nested folders. It is useful to know the maximum and average depths of the folder tree, not least because this may affect what would be appropriate visualisations of the tree structure. The top level folder in a structure has a depth of 0. A subfolder of this structure has a depth of 1; its subfolders have a depth of 2 and so on. One metric for assessing the relative depths of a tree is the maximum depth – the depth of the most deeply nested folder. However, trees are rarely uniformly deep, so two additional metrics can be used. The average depth is the mean depth of all folders in the tree. The average leaf depth is the mean depth of only the leaf folder (folders containing no subfolders).

Breadth. As well as varying in depth, tree structures can vary in width or breadth. If each folder tends to contain a high number of folders, the tree will be wider than if each folder has only a couple of subfolders. The more subfolders people create in each folder, the wider or ‘bushier’ their folder tree becomes. The average number of subfolders per folder is a measure of ‘bushiness’. The more files people store in each folder, the more ‘leafy’ their folder tree becomes. The average number of files per folder is a measure of ‘leafiness’. Higher leafiness indicates a denser tree. Another possible way of calculating breadth is to imagine that all subfolders were evenly spread amongst the folders and calculate how many on average would be in each folder. This metric is called branching factor.

Shortcuts. The Windows file system is a strictly hierarchical tree, which means that each file or folder can only exist at one location in the hierarchy. To facilitate more flexible navigation, the user can create shortcuts to files or folders from anywhere in their tree structure. This allows the tree to become a network, and allows more associations between items to be modelled. Because many applications create shortcuts on the Desktop itself, shortcuts here do not necessarily indicate the user is trying to create a more flexible document structure. Shortcuts within My Documents (or other document locations) or within subfolders of the Desktop are better indicators of attempts to use a more flexible structure.

Locations. The two standard locations provided by Windows for saving files are the Desktop and the user's personal My Documents folder. Users have the choice of using both of these, only one, or using other locations in addition to or instead of these. The number of separate locations in which documents are stored can indicate differences between folder structures, as can the number of files and folder in each top level location.

Duplication. Documents are sometimes duplicated in a file system for various reasons. A measure of the proportion of duplication can be calculated from the number of non-unique files divided by the total number of files. Duplication can be calculated separately for files and folders.

Top Level Proportion. Documents are often left unfiled in the top level, for instance, on the Desktop or directly in the My Documents folder. The proportion of files in the top level of the hierarchy provides useful information about this tendency. The proportion of folders at the top level of the tree also indicates how top-heavy the tree is.

File types. Users will vary in the types of files they work with most frequently. The document type of each file can be established. The number of file types employed and the number of files of each type in use can be compared.

The metrics discussed above can be derived automatically from the data. In addition, analyses of the file names and folder names can also be performed. Some of these can be automated (e.g., finding all dates used in folder names), however some of them will rely on semantic analysis.

4.3.2 Data Cleansing

In many cases, it was necessary to 'prune' the tree of data collected before commencing analysis. This occurred for many reasons, including the specific request of the participant to omit analysis of private or confidential files and folders. Another reason was that some participants had Microsoft Visual Studio installed on their machines. This software stores all its project files by default in a subdirectory of My Documents, although these are code and project files and executables, rather than documents. Therefore, this subdirectory (if it existed) was pruned from the analysis (after first verifying that it contained only the expected code and project files).

4.3.3 Analysis Software

Because the software to extract the file system data was custom written for this research, the software to interpret, display and analyse the data captured needed to be written as well. An Information Structure Analyser (ISA) application was written in Microsoft Visual Basic 2005.

Information Structure Analyser													
Summary	Participants	Analyse Structures	File Types	File Types	Folder Names	Graphs	File Type Occupancy	File Names					
Metric	...A...	...B...	...C...	...D...	...E...	...F...	...G...	...H...	...I...	...J...	...101...	...105...	...106...
Files	3790	20327	3793	1545	55314	3021	3614	372	3861	13380		524	170
Folders	228	2710	854	211	6614	415	609	39	191	616	0	46	22
FilesUnique	3345	4801	2891	1263	27359	1933	2585	347	2680	6449	0	463	162
FoldersUnique	199	493	491	131	2243	276	298	39	150	587	0	41	22
FoldersLeaf	170	1402	568	142	4384	289	339	32	153	534	0	30	15
FoldersEmpty	31	539	50	7	368	9	72	0	9	3	0	10	10
DepthAvg	2.27	6.43	6.13	3.77	5.9	6.02	4.26	2.21	2.15	2.97		1.78	1.64
DepthMax	6	16	11	8	12	12	8	4	5	7		4	3
DepthLeafAvg	2.4	6.83	6.38	4.04	6.03	6.36	4.52	2.34	2.17	3.06		1.93	1.87
FolderChildrenAvg	3.9	2.07	2.98	3.03	2.97	3.28	2.25	5.29	4.97	7.49		2.75	2.86
FolderChildrenStdev	5.26	2.43	4.02	3.41	4.28	5.26	2.5	4.64	13.11	14.89		3.77	2.54
FolderChildrenMax	32	26	58	20	67	44	28	13	82	111		16	8
FileChildrenAvg	16.62	7.5	4.44	7.32	8.36	7.28	5.93	9.54	20.21	21.72		11.39	7.73
FileChildrenStdev	61.84	16.46	6.26	7.69	21	16.92	11.06	9.11	31.46	25.37		22.49	14.88
FileChildrenMax	853	257	109	63	417	182	134	47	310	365		115	45
Locations	2	3	2	2	2	2	2	2	2	2	0	2	2
Shortcuts	35	4	9	5	24	36	12	4	10	14	0	6	10
DesktopShortcuts	35	4	4	2	10	14	10	3	8	12	0	2	8
DocumentShortcuts	0	0	5	3	14	22	2	1	2	2	0	4	2
OtherShortcuts	0	0	0	0	0	0	0	0	0	0	0	0	0
DesktopFolders	156	726	1	206	5710	3	1	1	79	1	0	1	1
DocumentFolders	72	33	853	5	904	412	369	38	112	615	0	45	21
OtherFolders	0	1951	0	0	0	0	239	0	0	0	0	0	0
DesktopFiles	3051	3802	25	1507	42044	35	14	6	1596	15	0	4	42
DocumentFiles	739	282	3768	38	13270	2986	2735	366	2265	13365	0	520	128
OtherFiles	0	16243	0	0	0	0	865	0	0	0	0	0	0
FileDuplication	11.74	76.38	23.78	18.25	50.54	36.01	28.47	6.72	30.59	51.8	0	11.64	4.71
FolderDuplication	12.72	81.81	42.51	37.91	66.09	33.49	51.07	0	21.47	4.71	0	10.87	0
DesktopFolderDuplication	14.1	69.15	0	38.83	68.13	0	0	0	51.9	0	0	0	0
DocumentFolderDuplication	2.78	6.06	42.56	0	34.62	33.74	48.24	0	0	4.72	0	11.11	0
OtherFolderDuplication	0	86.57	0	0	0	0	50.63	0	0	0	0	0	0
DesktopFileDuplication	11.9	55.05	0	18.51	50.58	0	0	0	67.54	0	0	0	0
DocumentFileDuplication	5.28	5.67	23.94	2.63	41.02	36.4	24.5	6.83	4.15	51.85	0	11.73	5.47
OtherFileDuplication	0	81.46	0	0	0	0	24.62	0	0	0	0	0	0

Figure 25: Information Structure Analyser software interface

4.3.4 Snapshot Results

Figure 26 below shows the extreme variability in the number of files in each person's snapshot. Harriet had the least with only 372 files, while Edward had 55,314.

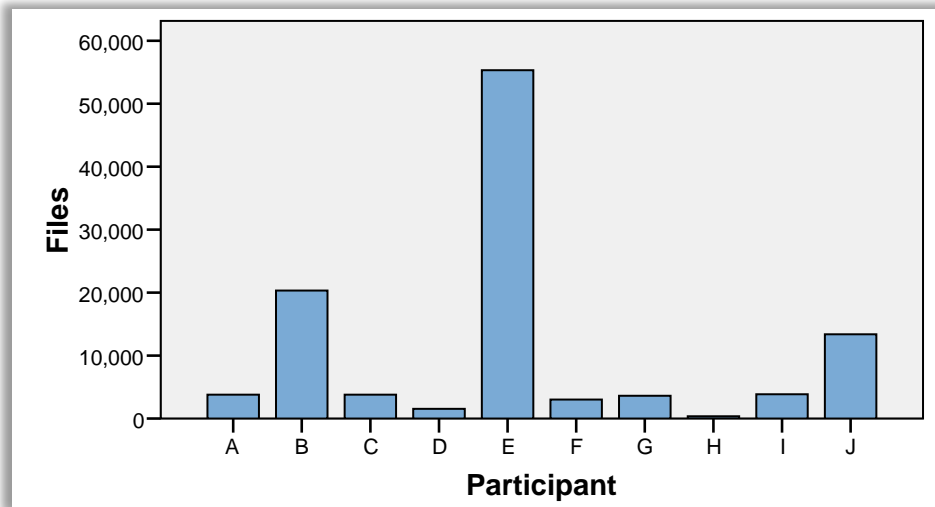


Figure 26: Bar graph showing number of files for each participant

The number of folders also varied significantly, as Figure 27 below shows.

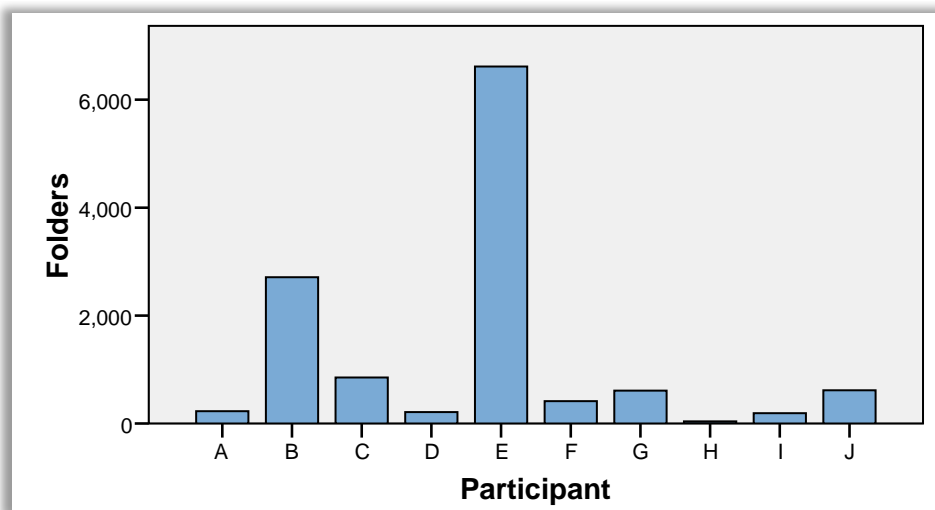


Figure 27: Bar graph showing number of folders for each participant

From a visual inspection, there is clearly a relationship between the number of files a person has and the number of folders they have. The correlation coefficient is .982 (sig=0.000).

Some participants had a significant number of duplicated files in their snapshots. Duplication was assessed as files having exactly the same name as files existing elsewhere in the system. As Figure 28 shows below, Brett, Edward and Jack have significant amounts of duplication. Approximately 50% of Edward's and Jack's files are duplicates, and for Brett, more than 70% of his files are duplicates.

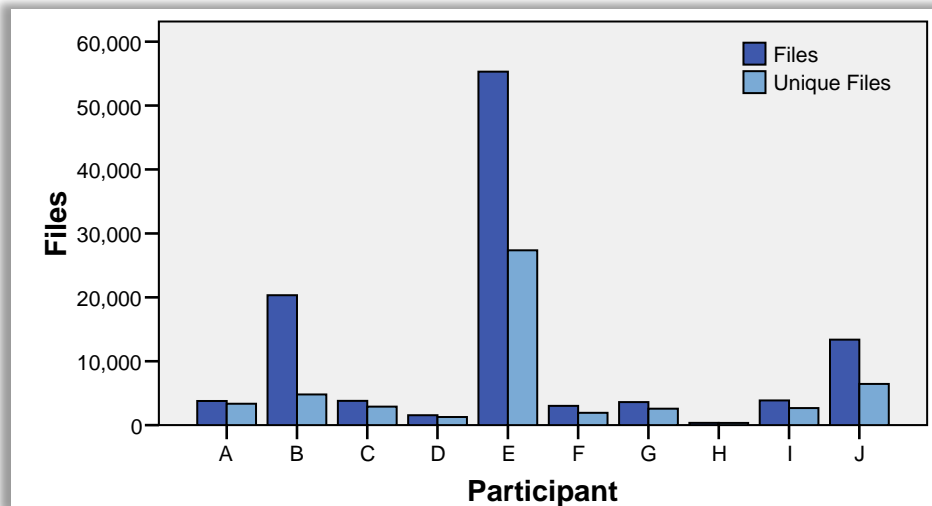


Figure 28: Bar graph showing unique files and total files for each participant

Some participants had empty folders in their snapshots. As **Figure 29** below shows, Brett and Edward both had a small proportion of their folders being empty. Most other participants had a very small number of empty folders.

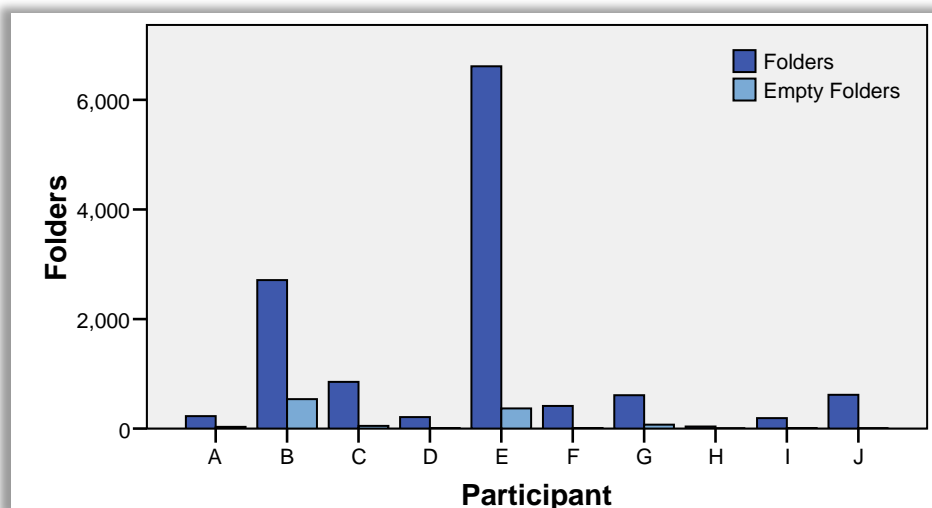


Figure 29: Bar graph showing empty folders and total folders for each participant

The depth of a participant's file system is the number of levels deep their folders are nested. There are two ways to look at this – (1) the maximum level of nesting they have anywhere in their system, or (2) the average depth of each folder's depth. **Figure 30** shows both of these values for each participant. Average depths tend to be between 2 and 6 levels. The deepest folders in a participant's folder system tend to be between about 5 and 12 levels deep. As you'd expect, there is a close relationship between the average and maximum depth values.

The participant's depth self-assessments were very accurate, with Ina, Harriet, Alex and Jack all stating that they tend to have shallow structures. These four participants had the lowest average depths.

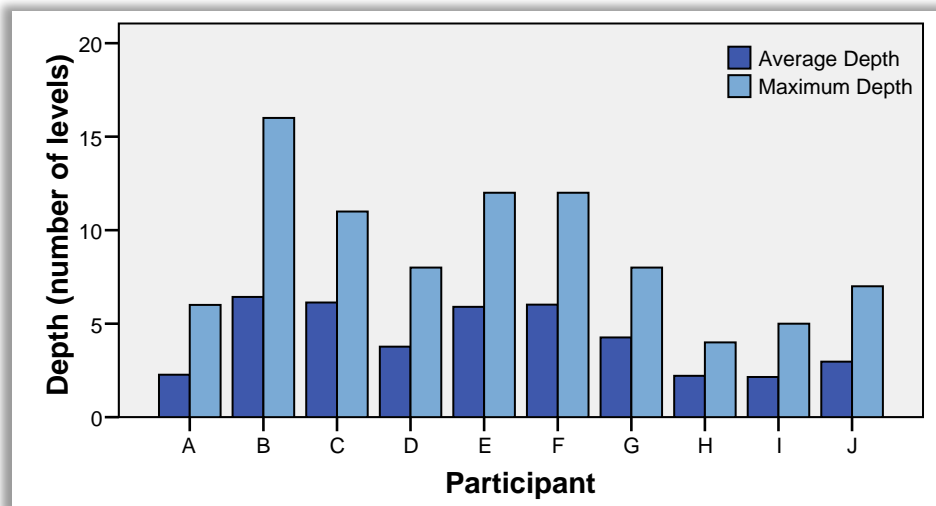


Figure 30: Bar graph showing average and maximum depth for each participant

The width of a person's folder structure depends on how many subfolders they have on average in each folder. **Figure 31** shows the average number of subfolders each participant keeps in their folders. These averages largely fall between 2 and 6, except for Jack, who averages 7.5 subfolders per folder.

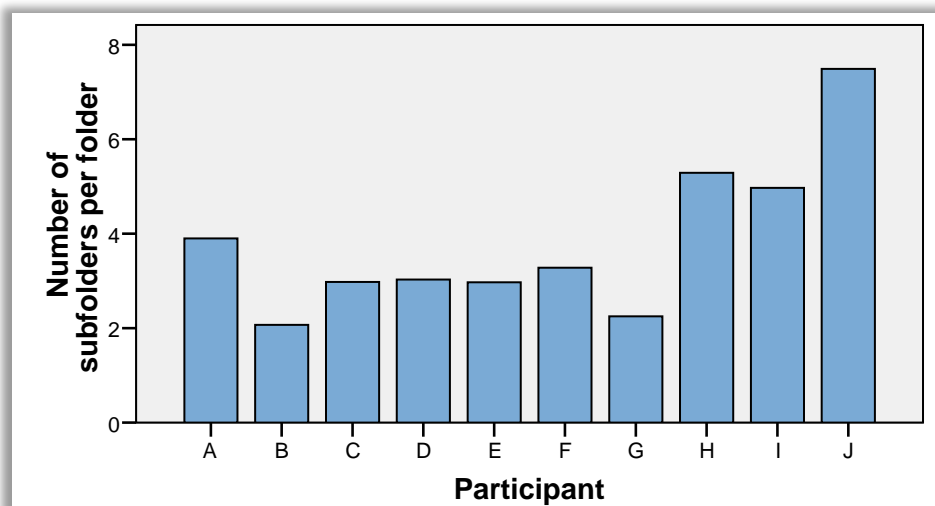


Figure 31: Bar graph showing bushiness - average number of subfolders per folder

While the average number of subfolders per folder is very useful, it doesn't tell you how evenly distributed the subfolders are. Someone with a very evenly distributed tree will have a low standard deviation of subfolder counts. A very unbalanced tree will have a high standard deviation of subfolder counts. **Figure 32** below shows the standard deviation of the count of subfolders per folder. Most of the players have a standard deviation below 5, but Ina and Jack have standard deviations of 13.1 and 14.9 respectively.

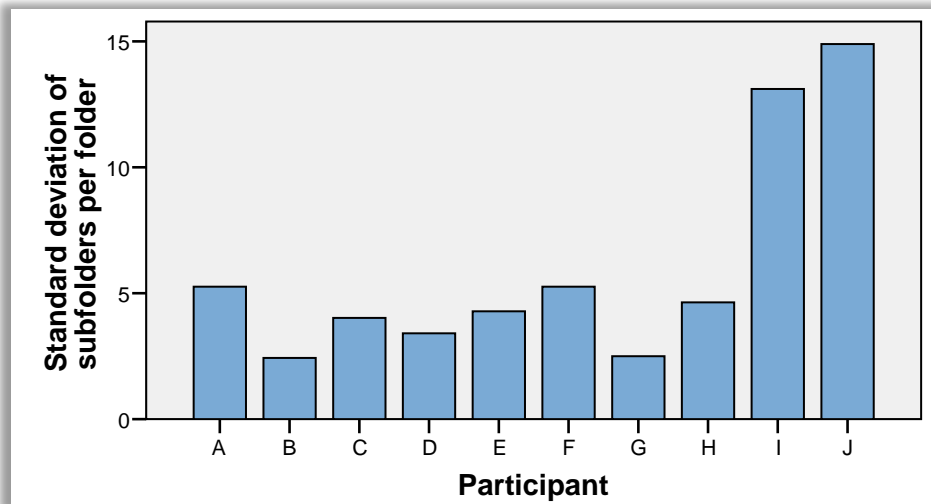


Figure 32: Bar graph showing standard deviation of subfolders per folder

Part of the reason for this can be seen if we look at the maximum number of subfolders each participant has. This is the number of subfolders in each participant's largest folder. As **Figure 33** shows, Ina and Jack have a very high maximum subfolder count.

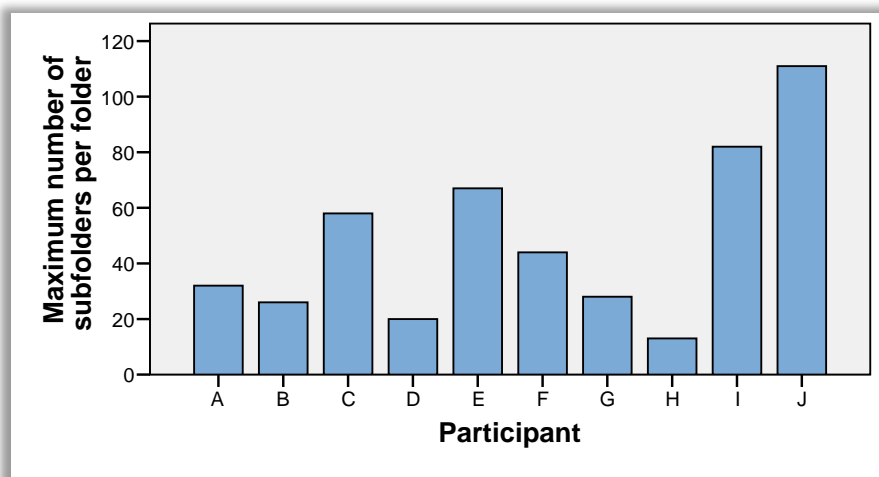


Figure 33: Bar graph showing maximum number of subfolders per folder

Branching Factor is a value that can be calculated to describe the shape of a tree. It takes the maximum depth and the total number of folders and calculates how many subfolders there would be per folder if they were all completely evenly distributed to the maximum depth. **Figure 34** shows the branching factor for each participant's snapshot. The branching factors range from 1.5 to 2.6, with Ina, Jack and Alex having the bushiest file systems.

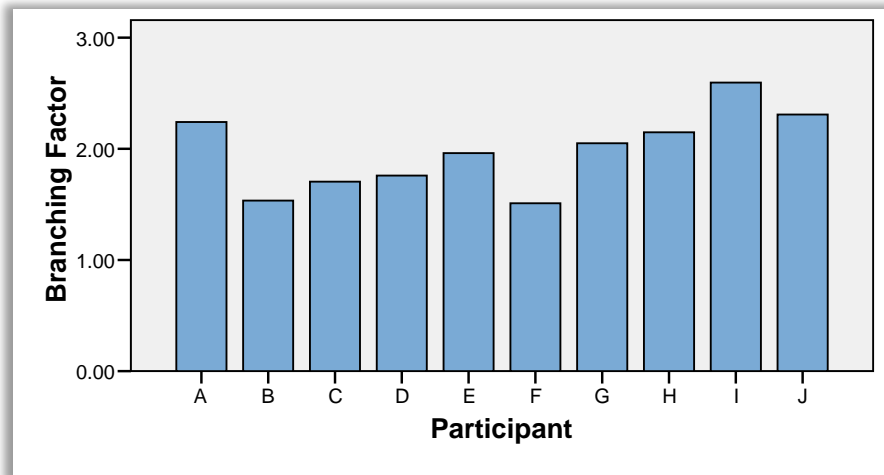


Figure 34: Bar graph showing branching factor for each participant

In a similar way to measuring bushiness of the folder tree using folder counts, we can measure the 'leafiness' of the folder tree using file counts. As **Figure 35** shows, the average number of files per folder ranges from 4.4 to 21.7.

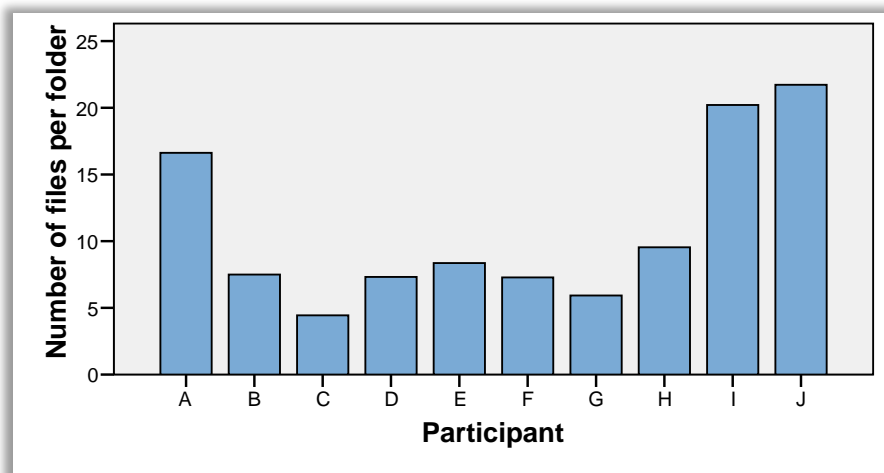


Figure 35: Bar graph showing leafiness - average number of files per folder

Again, to gauge the variability in the distribution of files, we can examine the standard deviation of files per folder. **Figure 36** shows that Alex has the highest standard deviation, a value of 61.8 compared with his average of 16.6. Ina and Jack also have higher standard deviations. This is a similar pattern to the folders, where participants with larger averages also have larger variability.

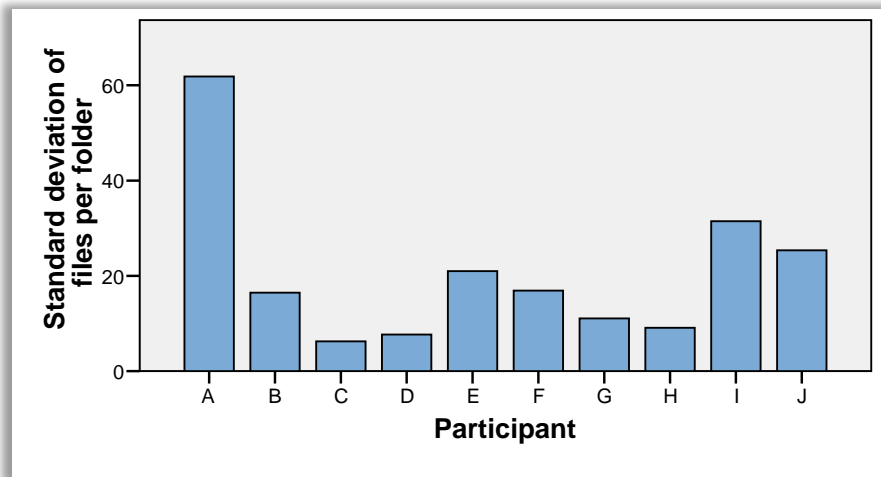


Figure 36: Bar graph showing standard deviation of files per folder

As with the folder, part of the explanation for the large variance can be seen in the maximum number of files per folder. Alex has a folder with 853 files in it, compared with his mean of 16.6.

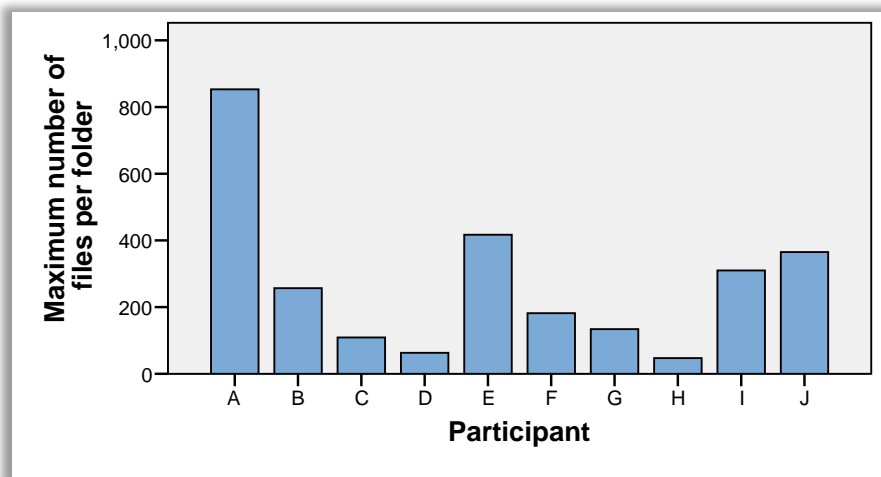


Figure 37: Bar graph showing maximum number of files per folder

Figure 38 shows the total number of shortcuts each participant has in the snapshot. Shortcuts are not particularly highly used, with the total number of shortcuts ranging from 4 to 36. If you compare this to the total number of files, the number of shortcuts is almost negligible, suggesting that the participants are largely using the file system as a strictly hierarchical tree.

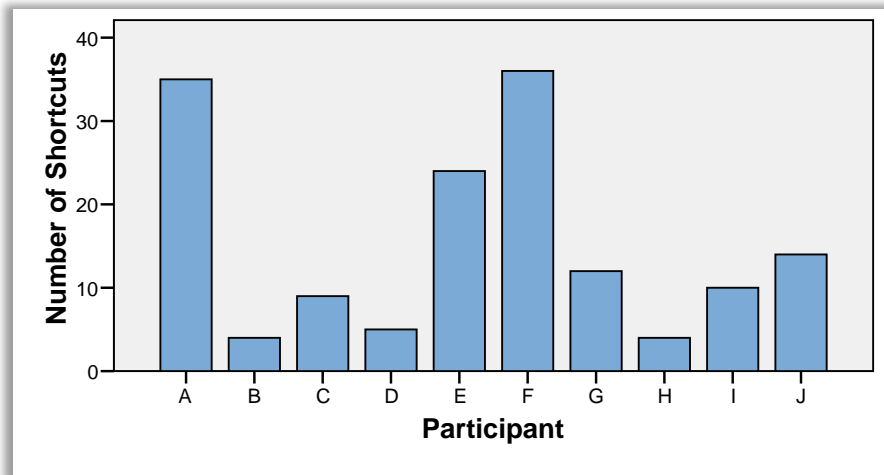


Figure 38: Bar graph showing number of shortcuts

Some applications automatically create shortcuts to their executable files and place them on the Desktop. These are not created by the participants themselves, and therefore don't really represent any document management activity. Because the snapshot did not capture the contents of any files (including the shortcut files) it is not possible to determine whether the shortcuts are pointing at executables or documents. However, we can examine the locations of the shortcuts. Shortcuts on the Desktop are likely to be application created, however shortcuts in the My Documents folder or its subdirectories are most likely to represent document management activity. **Figure 39** shows the total number of shortcuts in either the Desktop (or subfolders) or in the My Documents folder (or subfolders). Only Edward and Frank seem to have significant number of shortcuts in their My Documents folder. All other participants except Alex and Brett have small numbers of shortcuts in this location. Alex has all of his 35 shortcuts on the Desktop.

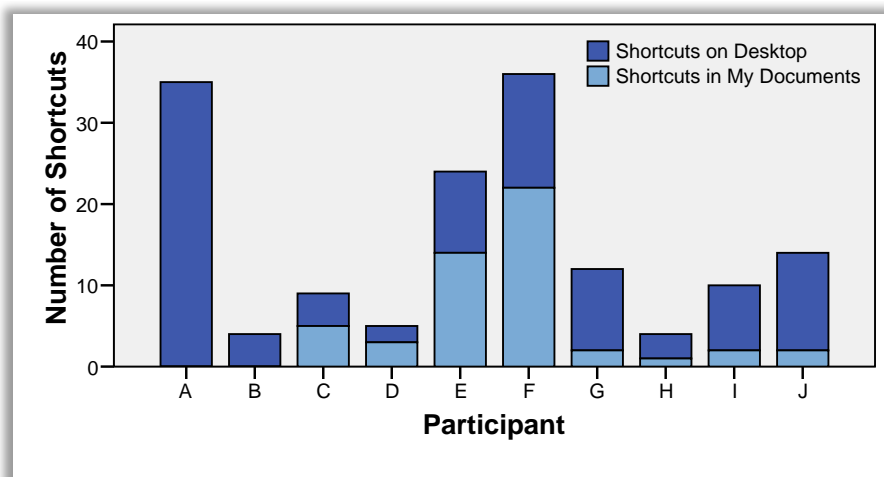


Figure 39: Bar graph showing the number of shortcuts in each location

We can also examine how the folders and files themselves are segmented between the Desktop and My Documents. Some participants stored their personal documents in other locations such as other directories on C drive, on network drives or external memory sticks. These have been grouped together

as Other Locations. **Figure 40** shows the total number of folders in each location. There is considerable variation in where the participants store their files. Alex and Edward have the majority of their files on the Desktop with the remainder in subfolders of My Documents. Brett has the majority in some other location with a sizeable minority of documents on the Desktop. Candice, Fred, Harriet and Jack have their documents entirely within My Documents and its subfolders. Damien has all of his documents on the Desktop and its subfolders. Gail has her documents split between My Documents and other locations.

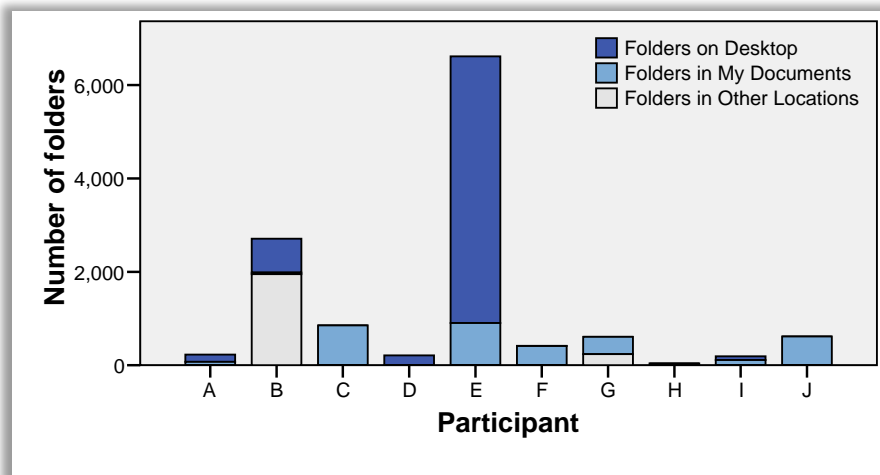


Figure 40: Bar graph showing the number of folders in each location

The distribution of files across locations is fairly similar to the distribution of folders, as **Figure 41** shows.

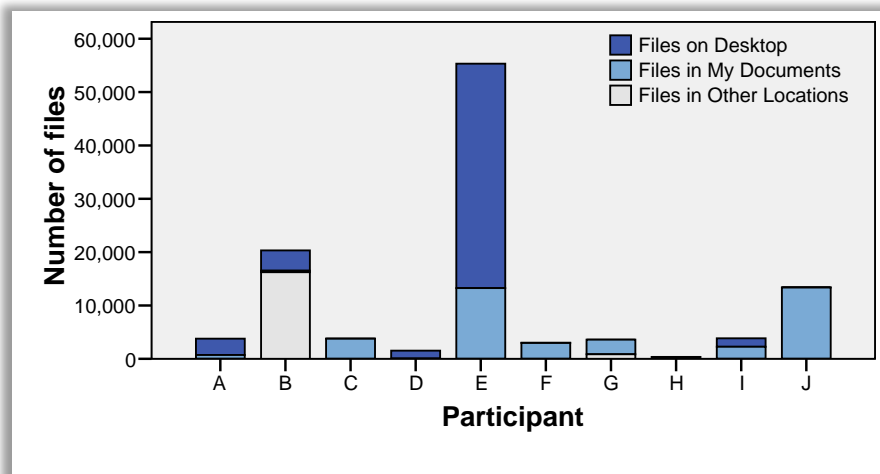


Figure 41: Bar graph showing the number of files in each location

Figure 42 shows the percentage of duplication of both files and folders in each participant's snapshot. Levels of file and folder duplication tend to be similar for each participant. This would suggest that the duplication arises because of the duplication of entire subfolders full of files, rather than isolated duplication of files. The exception to this is Harriet (and possibly Jack). Harriet has no folder duplication at all, and only a small proportion of file duplication. Jack has significantly higher file

duplication than folder duplication. The correlation coefficient between the proportions of file and folder duplication is 0.644 (sig=0.45).

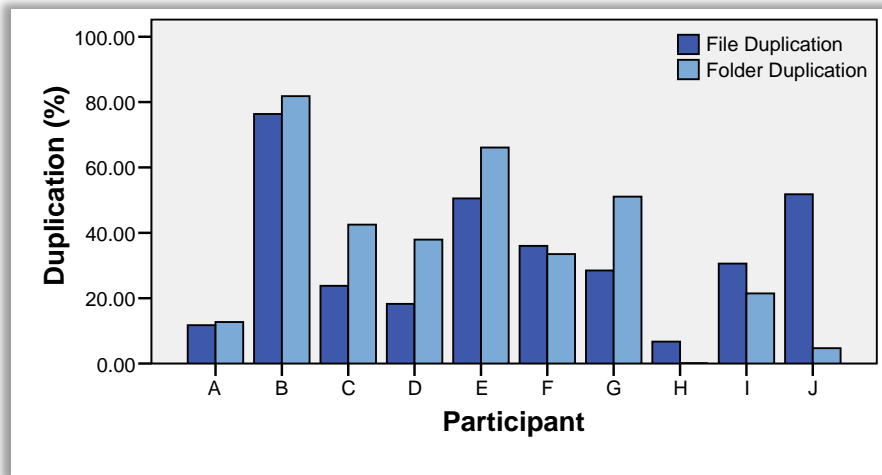


Figure 42: Bar graph showing the percentage of file and folder duplication

Figure 43 shows the percentage of folders that are in the top level of the folder structure. Folders in the top level are on the Desktop, inside My Documents or inside whatever other folder or device is used as a primary storage location. It is notable here that Ina and Alex have the highest proportions of folders in these locations. For both, this is a result of their shallow and broad hierarchies.

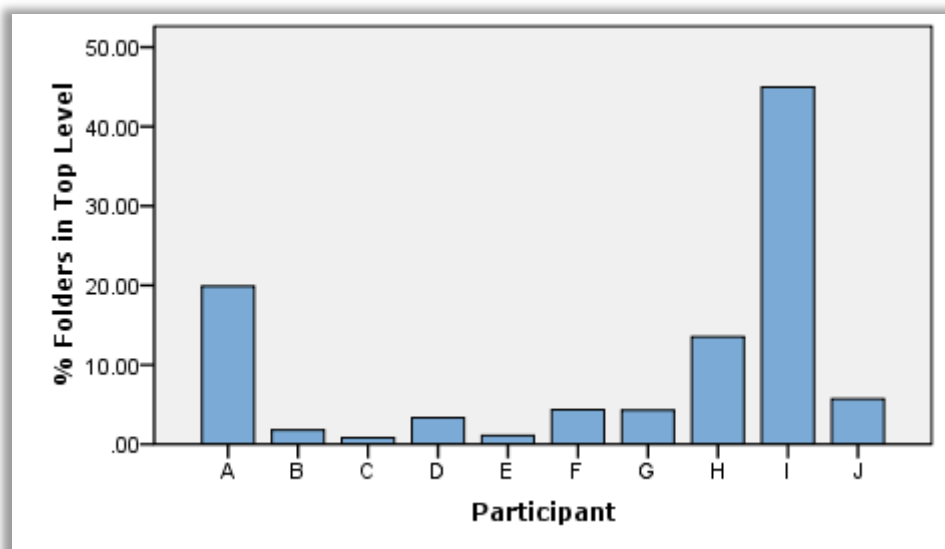


Figure 43: Bar graph showing percentage of folders in top level

The percentage of files in the top level of the structure is shown in **Figure 44** below. Harriet and Alex again have the highest proportions in these locations. Harriet's percentage is exaggerated by the small size of her document collection – while she only has 26 top level files, she also only has 372 files in total. Alex's smaller proportion reflects 188 files at the top level of his file system.

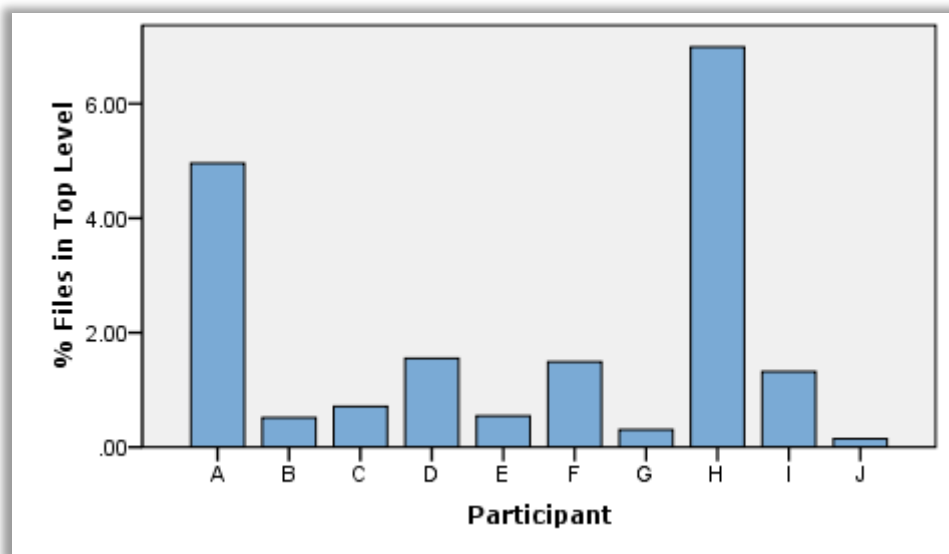


Figure 44: Bar graph showing percentage of files in top level

4.3.5 Folder Name Analysis

If there are commonalities in the types of folders people create, and the way they structure them, then there is scope for software to understand and support these processes. However, if folder naming is truly idiosyncratic, then there is less scope for automated support of folder creation and management. A folder name analysis was conducted for Alex, Brett, Candice, Damien, Edward and Frank.

The list of the folder names each participant has used was extracted from the file system snapshot information. The number of unique folder names was lower than the total number of folders, sometimes substantially. This was largely due to repetition of the same folder names in different places in the hierarchy, and in some cases due to wholesale duplication of folders structures.

Many of the folders were system generated, with the names assigned by the Operating System, or the software that created them. These include the default folders Desktop, My Documents, My Pictures and My Music etc. Many others were created when HTML files are saved, since a folder is automatically created to hold images and other resources with the same name as the file plus '_files'. Still others were created by installed applications. All of these that could be identified were eliminated from the list of folders, so the analysis was restricted to only the folders that the participants named themselves.

A list of unique folder names was created. Because the duplicate folder names were eliminated, the information about containment of folders was not displayed, and the coding was done from a simple list of unique folder names. These unique folder names were then inductively coded using a thematic open coding process. Every folder name used by every participant was assigned one more codes. The names assigned to the codes were continually examined to ensure that they accurately represented the material they were coding, and if necessary were changed to better represent their contents.

During the early part of the coding, new codes were added as needed, and sometimes codes were merged when it was recognised that they were actually coding the same concept. Eight codes were generated from Alex (all except Security and Source). Source was added for Brett, and Security was added for Candice. Both of these codes were also used by other participants so were retained.

There were many folder names that could not be assigned to any of these categories, and these were classified as 'Unknown'. The percentage of folder names that could not be coded ranged from 3% to 36%, and averaged 18%. This is to be expected, as folder and file naming is a personal and idiosyncratic affair, and some names may only have meaning to their creator in a particular context.

In some cases, information gained from the participants description of their folders was used to interpret some folder names that otherwise couldn't be easily classified. This helps ensure that the coding matches what the folder names meant to their creator as much as possible. However, this is necessarily an approximate process.

In order to check the extent to which the coding depended on the knowledge gained during the interviews, a second researcher coded Candice's folder structure using the established coding scheme. The percentage agreement between the coders was found to be 73%. After collaboration between the two researchers, this rose to 86%. This is surprisingly high given the subjective nature of the subject matter.

To assess patterns in the overall structure of the folders, an assessment was made based on the interview and from inspection of the file system snapshot. This is a necessarily imprecise estimate, as the folder structures of these participants had hundreds or thousands of folders, and it is common for different parts of the structure to be organised differently and to a different degree. To allow for this, the researcher also assigned a confidence assessment, of either low, medium or high, which indicated how pervasive the identified primary scheme seemed to be across the entire folder structure. If there seemed to be two equally pervasive schemes, both were identified.

The coding scheme inductively developed from the folder names is at a fairly coarse granularity. For instance, of the references to Time, some show a sequence (Week 1, Week 2), some indicate a relative age (Old, History, Archive), some indicate a particular year, and some indicate an exact date.

4.3.5.1 Folder Names

The folder names were coded to describe what type of information the folder name conveys about the contents (files and subfolders) of the folder. For instance, a folder named "Data Communications" tells you about the subject matter (topic) that can be expected in the folder, whereas a folder named "Lectures" tells you about the form and purpose (genre) of the contents.

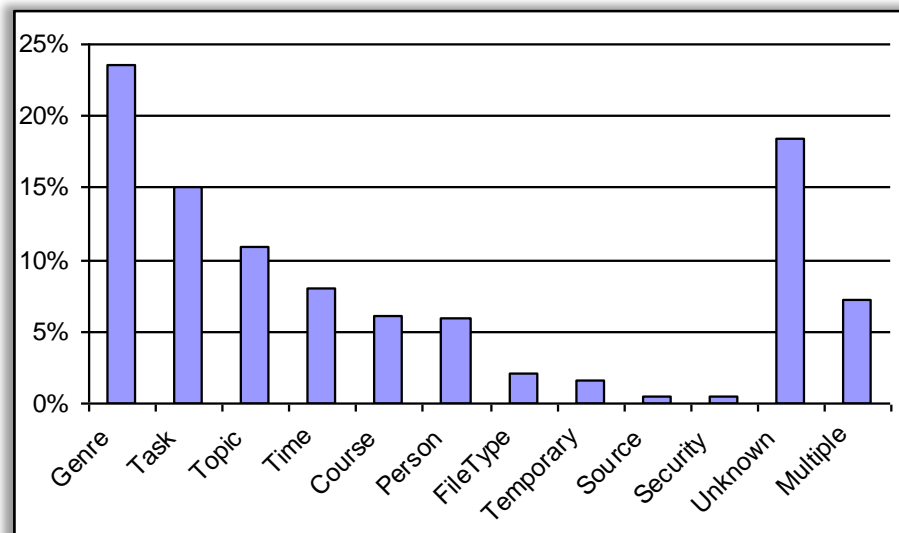


Figure 45: Average Proportion of Folders with each Code

Figure 45 shows the average proportion of folder names that was assigned to each code. The proportion of folders that were Unknown, and the proportion that was coded by multiple codes are also shown. Overall, the most frequently found types of folder name were Genre, Task, Topic and Time, followed closely by Course and Person.

Table 5 lists the codes that were derived from analysis of the folder names, with a definition and some examples of each code.

Table 5: Inductively generated codes

Code	Description and Examples
Genre	Indicates that the contents of the folder are a particular class or type of document, with a commonly recognized form and structure. Examples: Lecture Notes, Presentations, Timesheets, Budgets, Letters.
Task	Indicates that the contents of the folder are related to a task, project, event or some other type of activity. Examples: Assignment 5, Lec01, PhD, recruitment, evaluation, For DSS Presentation.
Course	Indicates that the contents of the folder are related to a specific course. (This is a special case of Task above) Examples: Database Systems, 222, INFOSYS 222
Topic	Indicates that the contents of the folder are all about a particular subject matter. Examples: Web development, Database Architectures, JavaScript
Time	Indicates that the contents of the folder are related to a particular time period, or have a time related aspect. Examples: 2005, 2003 SC, Old, History, Week12, Archive
Person	Indicates that the contents of the folder are related to a particular person, group or organization. Examples: Matthew, Audit Committee
File Type	Indicates that the contents of the folder are all a particular file format.

	Examples: zips, PowerPoints, Excel docs
Temp	Indicates that the name of the folder appears to have no intrinsic meaning and that little thought was given to assigning the name. Examples: foo, bar, fffff, asdfasdf, New Folder
Source	Indicates where the contents of the folder originated, either a location or person. Examples: From Brenda, From J Drive, Copy of R Drive
Security	Indicates that the contents of the folder are subject to particular security constraints or permission level. Examples: Personal, Confidential, Private

Overall, the most common type was **Genre**, and it is the one that deserves the most explanation. The genre of a document tells you what kind of document it is, something about its purpose and form. Orlikowski and Yates define it as a “distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form” (Orlikowski & Yates, 1994 p.543). For instance, knowing that a document is a newsletter gives us a different set of expectations as to what we can do with it than if we knew it was a journal article, a budget or a map. These distinctions are more to do with the purpose and form of the document than with the subject matter (topic) it is about.

Genre is something that is easily understandable to people but is rather difficult for a computer to understand. Assessment of document genre is not a completely objective classification, since different people can have different assessments of the genre of a document (Roussinov et al., 2001). In addition, people may deal with a vastly differing set of genres, depending on their job. Work is currently being done by Roussinov et al on automatic genre classification and using that to facilitate web searching (Roussinov et al., 2001), a line of investigation that these results suggest should be pursued further.

In this study, **Task** was defined quite broadly to include activities, projects and events, as well as more traditional tasks. Task often appeared in two quite distinct places in the hierarchy, either at the top or at the bottom. At the top, the tasks were very broadly defined, like Teaching and Research, and might more properly be thought of as roles. At the bottom, the tasks were more discrete, detailed activities, such as Tutorial 4. The concept of Tasks would be a good place to further investigate and refine the coding scheme with more participants.

The **Course** code is really just a specialized form of the Task code. Because all participants work at a University, courses figure prominently in their lives. It was decided to retain Course as a separate code, so that there would be flexibility to either separate it out or include it with Task as required.

Topic is something that can sometimes be determined by keywords in a file. A lot of research has been done on automatically categorizing documents into pre-established topic structures based on content analysis of the document itself (e.g. Apté, Damerau, & Weis, 1994). This is possibly the area where current automated assistance could be most useful.

Time was quite heavily used by most of the participants, although this is probably reflective of the fact that the study was conducted within a university. In an academic situation, the same course runs again and again and each instance of the same course needs to be distinguished from the others through some kind of time designation. It is entirely possible that in other situations that don't have regularly repeating activities, time would not be nearly so important. However, it is interesting to notice that the Software Developer also had moderate use of the time cue, and the two lecturers had almost no use of it at all. The use and importance of time in a wider setting is something that requires further work.

Most operating systems have time-stamping mechanisms that record when a file was created, accessed and modified, however, these are unlikely to be a substitute for the time dimension as observed here. As an example, consider a lab exercise that is handed out to the students. It is copied from a previous year's lab exercise, so the creation date is a couple of years ago. The accessed date is frequently changed by the backup process and other automated processes, but the modified date may give a reasonable indication of the age of the file. However, sometime software processes can interfere with this, such as an auto-save process saving the document while it is opened for printing. This could alter the modified date even if the file was not actually modified. Events like this mean that the modified date (as it is implemented by current file systems) cannot truly be trusted in order to locate the file in time, although it can provide a good starting point for assisting with automatic determination of date.

Person is a dimension of document that could be readily supplied by document management software. Already Microsoft Office documents include an Author attribute that is automatically filled in based on the login name of the user when the document was created. Mechanisms similar to this could be used to supply this attribute value.

File Type is an interesting code, since some researchers have suggested that the segmentation by file type is an artificial distinction that has limited relevance for document management and retrieval (Bergman et al., 2003; Boardman, Sasse, & Spence, 2003). Not only did file type appear in the way the folders are named, but during the interviews, all subjects reported sorting their documents by file type or searching on file type in order to quickly locate documents of a certain type. This would tend to suggest that file type is a necessary cue to allow people to distinguish and retrieve documents. However, it is possible that file type could be being used as a proxy for genre, since genre information is not available in current file systems. For instance, sorting by file type would allow you to easily distinguish between a presentation (typically a PowerPoint file), a journal article (commonly stored as PDF), and a budget (likely to be an Excel file). Although this doesn't have very much power to discriminate between documents, it nonetheless might be useful for that purpose in the absence of

genre information. These findings would suggest that more investigation needs to be done on the usefulness of file type before it is dismissed in the design of future document management systems.

Source is another code that is very interesting, although not commonly used. The folders coded as Source also included two that were actually specifying a destination, but since this only occurred twice and for one participant only, it was not coded separately. However, source was also sometimes implied in folders named for people, such as 'Annie's lectures', and 'Jim's'. Whether a document arrives as an email attachment, is downloaded from somewhere or copied from another location, the document management software should be able to detect where it came from and automatically store that information.

Since email is now the primary form of document exchange (Ducheneaut & Bellotti, 2001), most documents that were not created by the owner probably arrived as email attachments. This gives potential for even more information to be stored, such as exactly who they came from and when. Ideally, this information should be accessible when managing documents, and not solely in the email system.

Security was the least frequently used of the codes, but was encountered in three of the six participants. This designation of certain things as private, confidential or shared is something that could be easily supported by document management software. Since it seems that security designations are relatively seldom used, it would probably be appropriate for it to default to a 'Normal' setting, and which could be changed by the user when required.

The **Temp** folders were an interesting group, although they only appeared in three of the participants. More investigation needs to be done on why these folders were created, and what breakdown in the process is causing them to appear. It would also be interesting to know how long they last, and whether they are eventually given a more meaningful name, or deleted altogether.

Table 6 shows the percentage of each participant's folders that were classified using each Code.

Table 6: Proportion of Folders Coded with each Code

Code	A	B	C	D	E	F	Avg.
Genre	12%	20%	29%	32%	32%	13%	24%
Task	11%	17%	4%	3%	1%	55%	15%
Topic	7%	5%	7%	7%	23%	15%	11%
Time	8%	6%	22%	9%	1%		8%
Course	2%	2%	10%	20%	1%	3%	6%
Person	8%	1%	9%	16%	1%		6%
Temp	8%	1%	<1%				2%

File Type	4%	1%	1%		3%	3%	2%
Source		<1%	1%		2%		1%
Security			<1%	2%	1%		0%
Multiple	2%	25%	3%	7%	1%	8%	8%
Unknown	36%	20%	13%	3%	35%	4%	18%

This table shows the percentages after system-created folders have been excluded from the analysis. Due to rounding for display purposes, the totals in each column may not add up to exactly 100%.

For all participants except Frank, Genre was the most common type of folder name encountered. For Frank, it was outweighed only by Task. Folders of type Temp, Source, and Security were only employed by three of the six participants in this study. Frank only exhibited use of five of the 10 codes, while Candice showed use of all of them.

Figure 46 shows radial graphs of how each of the six participants different with respect to the proportions on their documents that were classified according to the top four codes. For the purposes of this graph, Course has been included with Task. This gives a graphical view of the profile of the top four codes for each participant.

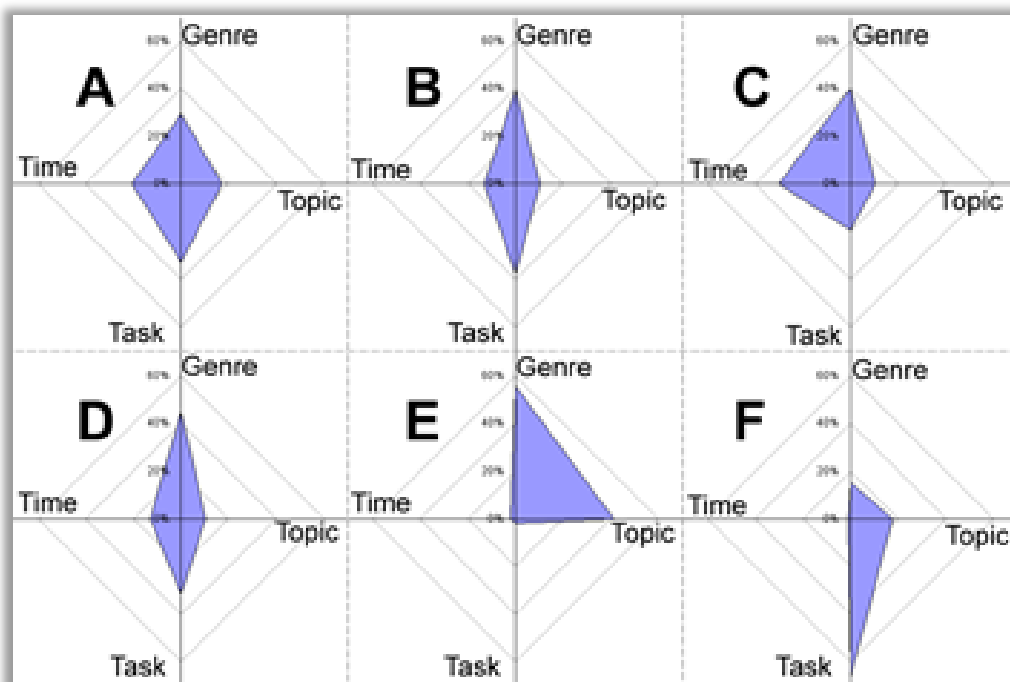


Figure 46: Radial Graphs showing profile of the top four codes

Alex, Brett, Candice and Damien all have reasonably similar profiles. Alex has an approximately equal proportion of each of the four codes, with a slight tendency to more Task-based folder names. Brett and Damien have a very similar distribution, with a tendency towards more Task-based folders and

fewer Time-based folders, whereas Candice tends to the opposite, with more Time based and fewer task based. Brett, Candice and Damien are all Course Managers, performing essentially the same duties, so it is perhaps expected that their folder name profiles are similar to each other.

However, Edward and Frank both have the same position (both are Lecturers, with similar teaching and research responsibilities), but their folder name profiles show quite different tendencies. Edward tends to have mainly Genre and Topic oriented folder names, with very little use of the Task or Time dimensions, whereas Frank tends to have overwhelmingly Task based folder names, with some Genre and Topic but no Time-based folders at all.

Multiple Coded Folder Names

In addition to the folders that were coded with a single type, some folder names were assigned multiple codes. This happened when the folder name was made up of multiple parts, each which could be classified differently. Some examples of common multi-code folder names are given here. If a folder could be classified as more than one name, it was given a primary code and a secondary code.

Not all possible combinations of codes were actually present in the folder names of the participants. **Table 7** shows examples of combinations that were actually used.

Table 7: Examples of Multiple-Coded Folder Names

Folder Name	Primary Code	Secondary Code
2004 Programming	Time	Topic
INFOSYS 222 2005AC	Course	Time
Recruiting 2003	Task	Time
INFOSYS 222 Exams	Course	Genre
Jim's Timesheets	Person	Genre

Table 8 shows the frequency of different combinations of these multiple coded folder names. These are relative percentages of multiple-coded documents, not percentages of all documents. These percentages add up to more than 100% due to rounding for display purposes.

The most common combination of Time and Topic was largely due to Brett, who had 25% of his folders multiple coded. Brett has an ongoing project that he works on every day. Every time he finishes working on it, he makes a new copy of the entire folder containing the project material, and names it with the current date, as well as the topic that he most recently addressed in the project. These folders accounted for almost all of the Time and Topic coded folders, and also accounted for the high level of duplication in Brett's folders, since he creates a duplicate copy of the entire folder structure on a daily basis.

Table 8: Combinations of Multiple Coded Folder Names

Secondary →								
Primary ↓	Course	Genre	Person	Task	Time	Topic	File Type	Security
Course		6%		3%	14%	2%	1%	
Genre	1%		1%		3%		1%	1%
Person		4%			1%			
Task		6%			7%			
Time		1%				38%		
Topic		4%	1%		1%			

The next most common set of codes was the combination of Course and Time. This is unsurprising, since courses run over multiple semesters, and it is common at this University to refer to one particular instance of a course using the course code, year and semester designation.

Other combinations include the combination of Task and Time, commonly seen in names like Lab Week 4, Recruiting 2003; the combination of Course and Genre, as in INFOSYS 222 Tutorials; and Task and Genre, as in Lecture 5 Demos.

It is very interesting that only 8% of the folder names were multiple coded, and even that figure was very much inflated by Brett's naming practice. This doesn't necessarily mean that the folder names are single word, just that the folder names tend to only represent one of these dimensions. Thus, it seems that these multiple dimensions are applied to documents through combining several single-dimension folder names into a hierarchy, rather than by constructing multiple dimension folder names and using those.

One plausible reason for this is reuse. Especially in the university environment where courses repeat year after year, it is possible to copy a folder and all its subfolder and documents, and by changing one of the folder names higher up the tree, change the context of all the documents below it. Consider the example in **Figure 18** in **Chapter 2**. By simply copying the 2005 folder and naming the copy 2006, the user can change the context of all the documents and folders below it. It would be interesting to see whether the same findings occur in a situation where there was not such strong time based replication as in the university.

In addition to these folders that are dual coded, **Table 9** shows the combinations of triple coding that were found.

Table 9: Triple Coded Folder Names

Code	Example	Incidence
Course+Time+Genre	INFOSYS 222 2005 AC Tutorials	1%
Person+Topic+Genre	Yin's Modelling Tutorials	1%
Time+Genre+Source	OldExamsFromEmma	2%

This triple coding was relatively uncommon, accounting for less than 0.5% of the folder names overall, and less than 5% of the multiple-coded names, and only appearing in two of the participants' file systems. This is probably due to the fact that more flexibility is obtained from nesting single folder names than from encoding the same information into a single folder name.

For example, consider the case of Course+Time+Genre, and a folder called "INFOSYS 222 2005F Tutorials 222" representing the tutorials for a database course taught in the first semester of 2005. These three pieces of information can all be encoded into a single folder name, or the same information can be discerned by nesting three separate folder names. The following three hierarchies would all give the same information:

- INFOSYS 222 > 2005F > Tutorials
- Tutorials > INFOSYS 222 > 2005F
- 2005F > INFOSYS 222 > Tutorials

Having the separate folders allows other documents and folders to be placed into the intermediate levels, providing context to those without having to create additional folders.

4.3.5.2 Folder Hierarchy

Table 10 shows the primary organization scheme employed by each participant. The confidence column indicates how much confidence the researcher had in how rigorously this scheme was followed throughout the file system.

Of interest is the different primary organising schemes used by Brett, Candice and Damien. These three participants are all Course Managers whose profiles of folder types were all very similar to each other; however, their dominant organising schemes are all quite different. None of these people expressed significant dissatisfaction with their organising scheme, and all seems able to effectively use their structure to perform their jobs.

This would suggest that perhaps the order in which these folder types are combined is not particularly important. As noted before, if you place a document in the bottom level, the combination

of all the folders in the hierarchy above supply the required meta data to be able to distinguish the file from others, regardless of the order in which they were encountered.

Table 10: Primary Organisational Scheme

Participant	Scheme	Confidence
Alex	Time > [various]	Low
Brett	Time > Course > Task	Medium
Candice	Genre > Time	Medium
Damien	Task > Course > Time > Genre	High
Edward	Task > Time > Course > Genre or Task > Course > Time > Genre	High
Frank	Genre/ Task > [various]	Low

During the interview with Edward, he noticed that what he thought were two identical folder structures that he maintained in two different places, were actually different. In one place, the order was Task > Time > Course > Genre, in the other it was Task > Course > Time > Genre. Despite priding himself on keeping these two structures perfectly synchronized with each other, he'd never noticed that in fact the order of two of the primary dimensions was different. This would seem to confirm, that for this participant at least, it makes little difference which order these dimensions appear in. He commented that there was "nothing to choose between them".

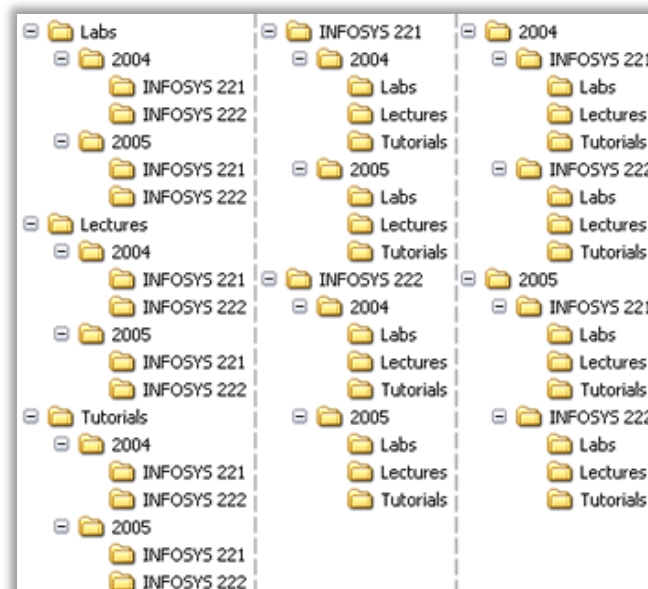


Figure 47: Three equivalent folder structures

Each of the three folder structures shown in **Figure 47** encodes the same information about the files and folders in the bottom level. All three of these structures are showing ways to assign three dimensions to a documents: a course dimension (2 options), a time dimension (2 options), and a genre

dimension (3 options). Since there is no natural subordination between these dimensions, any of these three structures will do the job. However, precisely because there is no natural subordination, all of them involve considerable duplication of information.

One possible way of overcoming this problem would be to consider these dimensions as separate facets of the document, and to allow the user to provide values for these facets separately. The document management system could then allow any of the views shown in **Figure 47** to be dynamically created, and to be changed on the fly by reordering the facets as needed. In addition, the facets could be used to filter the information displayed if required. Hearst et al (Hearst et al., 2002) have had success in using a faceted search system called Flamenco, which operated on a collection of landscape and architecture photographs. Although personal document collections have quite different characteristics to an image library, this technique seems a promising avenue of exploration for document management.

4.4 DOCUMENT MANAGEMENT STRATEGIES

Putting together the information from the interviews and the snapshots, the participants seem to loosely adopt one of three main strategies: piling as practiced by Alex, structuring as practiced Damien, Frank and Harriet, with the remainder adopting a filing strategy as necessary to ensure retrieval.

4.4.1 Piling

Alex doesn't really file his documents, he just lets them pile up on his Desktop until it is full and he dumps them into a folder. Because of this he has a fairly shallow and broad file system, with a higher proportion of documents at the top level of his structure. He doesn't really use the tree (since he doesn't really have many folders organised into a hierarchy) and is more likely to search for lost files if he can't find them using chronology. He is also likely to use the details view more often so he can sort items easily. He rarely creates items in advance and because of this lack of structure, considers himself relatively disorganised.

4.4.2 Filing

Most participants file documents into folders in order to retrieve them later. They split folders up if the number of documents grows so large that they cannot easily spot items within them anymore. They tend to create folders either during cleanups or just-in-time as they need to save a folder that doesn't fit an existing category. They do have a hierarchy, although it is moderately broad and not particularly deep. They are likely to have some files in the top level (pending cleanups), and quite a few folders as well. There is no particular preference for view, but they are much more likely to locate files by browsing their structures than searching. They would generally consider themselves to be relatively organised.

4.4.3 Structuring

Structuring filers such as Damien and Frank share many of the same behaviours as filers, but for them the creation of a structure serves a larger purpose than simply retrieval. It also provides them with an overview of the structure of their information, and for this reason they will often create folders in advance of having files to put in them, simply because seeing the conceptual categories is useful. They are likely to have fairly deep and narrow structures, and to have relatively few documents at the top level of their folder structures. They are more likely to browse through their structures although because there are so many places to look, will more readily search for older files. They tend to consider themselves very well organised.

4.5 CONCLUSION

This chapter has described the interviews and file system snapshot components of this research. In this study, 10 participants were interviewed about their document management practices and a snapshot of their file system was taken. This chapter describes the process of inductive thematic coding that was performed on the interview transcripts and presents a conceptual model of the resulting themes.

Participant's statements about their documents can be grouped into six main areas: Attitudes, Creating Files, Folder Structures, Finding, Retention and use of the Desktop and My Documents folders in Windows XP.

Participants generally felt that it is desirable to be "well organised", expressing either pride or guilt depending on whether they felt they had achieved a good level of organisation. Reorganisations were common, with some people performing them periodically (often on an annual or per-semester basis) and others continuously. Others thought they should reorganise and wished they could but lacked either the time or the motivation to do so. Some participants prefer to archive files that are no longer needed from their primary document collection, others prefer to leave them in place.

Participants commented favourably about the hierarchical nature of file management in Windows XP, generally indicating they found the hierarchy intuitive and a good match for their mental models. Several participants identified their own folder structures as shallow, having few levels of subfolders.

The idiosyncratic nature of document management was clear in the approach to file creation and folder naming. Document naming tends to be ad hoc and personal, with people trying to choose names they found meaningful to them, with little use of systematic naming conventions. People reported folders being named in various ways, with document genre, topic, time and course or project being common labels used. Individual differences were also apparent in methods of locating documents, with various combinations of searching and browsing employed according to personal style.

Participants mentioned the ability to group things into folders, and the ability to change views and sort items as being amongst the most useful features of the Windows XP's document management capabilities. The time taken to navigate using the tree was mentioned as one possible drawback, as was inability to replicate structures, and the lack of context provided by the file management tool. Most couldn't think of anything missing, with one request for more spatiality and one request for dynamic folders.

The file system snapshot showed a surprising diversity in the size of document collections being managed, ranging from a few hundred files to tens of thousands. Duplication was clearly a factor in this, with copies of files making up more than half of the total number of documents for some participants. File structures tend to be between 2 and 6 levels deep, and participants with shallow folder structures seem to be quite aware of that fact. Folders which contain other subfolders generally tend to have between 2 and 6 subfolders, although there was individual variation between participants. Individual variation also plays a part in the number of files kept in each folder, with seven of the participants averaging between 5 and 10 files in each folder, but three others averaging between 15 and 22. This would tend to confirm the idea that for some people there is a conscious or subconscious limit after which people seek to further categorise their information. Shortcuts were used infrequently and largely on the Desktop for quick access to applications or folders, indicating little inclination to turn the hierarchy into a network.

The analysis of folder names identified a number of different organising dimensions which provides a starting point for understanding how people structure their documents, with the four most common being document genre, task or project, topic and time. These dimensions can be combined in many different ways, since there is no "best" way to combine them into a hierarchy. Folder names with multiple codes accounted for less than 20% of folders, with multiple labels generally being applied through containment in the hierarchy. Most of the multiple coded folder names arise from combinations of topic and time. Others stem from the repeating nature of courses over time in an academic environment.

The combination of interview and file system snapshot yielded three broad strategies that people can adopt with respect to their documents: piling, filing and structuring. Users of the three strategies differ in the approach they take to work with and retrieve their information, and as a consequence, the structures they create also differ.

In the following chapter, this conceptual model will be used to develop a questionnaire that along with the file system snapshot can be administered to a larger number of participants in order to validate the conceptual model developed in this chapter.

Chapter 3 provided a rationale for using a survey research strategy to follow on from the interviews described in **Chapter 4**. The conceptual framework outlined in the previous chapter provides the basis for developing the survey questionnaire, and this questionnaire will be delivered in tandem with the file system snapshot that was also described in the previous chapter. The questionnaire and snapshot together allows the acquisition of empirical data from a larger number of participants than is feasible with field studies, and thus provides information about the frequency of the strategies and behaviours observed, and thus allowing the conceptual model to be validated.

Section 5.1 describes the design and implementation of the survey, **Section 5.2** then presents the results obtained from the questionnaire and **Section 5.3** presents the results from the file system snapshot. **Section 5.4** describes a classification model based on a selection of survey questions and snapshot metrics. A discussion of these results and how they reflect on the conceptual model is deferred to the following chapter.

5.1 SURVEY DESIGN

This section describes the design of the survey, including the design of the questions, and the design of the participant contacts and the questionnaire itself.

Ideally, an existing questionnaire should be used to investigate this topic (Bourque & Fielder, 1995), however, since no studies like this have been done in the past, there are no existing questionnaires that can be used. The next best thing to using an existing questionnaire is to adapt an existing questionnaire for the purpose. However, due to the specific nature of the knowledge required in this area, there are

not even existing questionnaires similar enough to be adapted and be useful for this research. Therefore, the only option is to generate new questions from scratch.

The conceptual framework described in the preceding chapter (shown in **Figure 22**) provides the basis for generating the survey questions. The survey questionnaire is designed to explore those concepts, as well as to uncover more quantitative information about the range of document management behaviour.

An initial set of questions was derived from the conceptual framework. For instance, on the theme of version identifiers, the following set of draft questions was generated:

- Do you use separate files for different versions of documents?
- Do you delete old versions of documents?
- Do you copy old version of documents into other folders?
- Do you use version indicators on your file names?
 - Do you use numbers as version indicators?
 - Do you use dates as version indicators?
 - Do you use people's names as version indicators?
 - Do you use descriptions as version indicators?
- Do you ever have trouble finding the latest version of a document?
- Do you initiate versioning schemes when collaborating with others?
- Do you end up editing two versions of the same document independently?
- How do you reconcile two versions of a document that have been edited independently?
- Do you use the revision/track changes feature in Microsoft Office applications?
- How big a problem are versioning issues?
- What could be done to better support versioning?

The initial survey questions were derived to elicit information covering the range of concepts that were generated from the interviews. This generated over possible 300 questions, not including demographic questions. The next step was to work out linking between questions. For instance, if a participant doesn't keep separate files for versions, then none of the other versioning questions are relevant to them and should be skipped.

At this point there were still over 200 questions to be included in the survey. The longer a questionnaire is, the less likely people will be to fill it out, and the less accurate their answers will be, especially towards the end of the survey (Deutskens, Ruyter, Wetzels, & Oosterveld, 2004). Although there is no definite limit on how long is too long, research has indicated that a survey which can be completed in 15 minutes is a good length (Porter, 2004). As the participants were also being asked to run the file system snapshot software, the number of questions had to be reduced to a significantly more manageable level. Although this meant that many potentially useful questions had to be discarded, the question set was cut down to 85 questions. All major sections of the conceptual framework were still represented in the question set.

The first step in refining the questions was to decide what form of question it would take. While open-ended questions can provide richer data, they are considerably more difficult to code and analyse, and participants are less likely to complete a survey that requires a large amount of writing. For this reason, most questions were not open ended. Most of the questions took the form of a yes/no choice, a frequency assessment or an agreement assessment, with many questions having multiple phrasing possibilities. While many of the questions were categorical questions, some were ordinal questions, using Likert scales to allow the participants to rank importance, agreement, or frequency. For these, a five point scale was used, which provides discrimination without overwhelming the user with fine distinctions. As recommended by Fink, the negative end of the scale was presented first (Fink, 1995a). The question itself was phrased to use both ends of the scale in order to avoid biasing the respondent and to indicate that either set of options is acceptable. For instance, Question 2 asks 'How much do you agree or disagree with ...', rather than simply 'How much do you agree with'.

The predominance of closed questions makes analysis easier, but these questions are more difficult to design since the responses must be known in advance. If the range of responses offered does not match the responses that the participants want to give, the question won't give a meaningful result (Fink, 1995a). For this reason, the options for each question were designed to include all the behaviours and opinions observed in the interviews. In some questions, respondents are being asked about particular features in the Windows file system and the set of responses is naturally bound by what the software offers. An example of this is the question which asks respondent what is displayed on their Desktop. There are only six possible response options for this question because Windows XP only offers six possibilities. As well as having a limited set of possible options, the question is easy to answer and applicable to everyone and so provides a good icebreaker question for the start of the main section of the survey.

Additionally, the questions were progressively refined to use language that would be more easily understandable to someone with little or no computer knowledge, follow the guidelines for writing good survey questions. In particular, the questions were made as concrete and unambiguous as possible, using clear, conventional language with complete sentences (Fink, 1995a). The questions were short and specific, logically grouped and (where possible) progressed from easier questions to more difficult ones (Bourque & Fielder, 1995).

Particular care was taken to use neutral language, and especially to avoid emotionally charged language. Phrases like badly organised and disorganised have negative connotations and may cause the respondent to resent having to characterise themselves that way.

On almost any given element of document management behaviour, most of the interview participants reported using a range of different techniques, some very frequently and some very

occasionally. For instance, many of the participants who use version identifiers could probably find examples of files when they have used version numbers, files where they have used dates, files where they have used people's names and files where they added a description. In many of the questions in the survey, if the question was phrased in the form of 'have you ever ...' most respondents could probably choose many or all of the options presented. For this reason, most questions were phrased with the word usually, or the phrase most often. Thus, simply asking if they used a certain technique would quite likely result in most people saying they did and not really provide any information about whether it was something they commonly did or something that was very rare. For this reason most questions were phrased to require a frequency assessment, and the word 'usually' is also used frequently in the survey. The aim is to capture the predominant personal document management behaviour of the respondent, not to catalogue every possible behaviour they have ever engaged in. Almost every participant is likely to have used other techniques or approaches at some time in their past, particularly when new to a system.

In other questions, particularly where respondents are being asked about their reasons for certain practices, a 'residual Other' has been used to increase the flexibility of the answer categories (Bourque & Fielder, 1995). The 'residual Other' technique involves providing an additional category called Other, and allowing the participants to provide an open-ended response. An example of this is the version identifier question:

- How do you usually distinguish between the different versions?
 - I put version numbers in the file names
 - I put dates in the file names
 - I put descriptions in the file names
 - I put the files into another folder
 - Other: _____

The first four options were observed in the interviews. The final option is present in case some respondents did something altogether different, or a combination of the above. They can enter any description they choose

When analysing the responses to this type of question, the response given in the freeform text area needs to be carefully analysed to determine whether or not it should properly belong in one of the existing categories, or whether a new category should be created for that response. A large number of responses in the 'Other' category tend to indicate that the response set for the question was not appropriate.

At this point in the survey development, the questions were categorised, so that related questions could be presented to the respondents together. Fourteen categories were created, plus a section for

Demographics, in order to avoid having any section containing a large number of questions that respondents might find off-putting.

The order in which questions are answered can be important in ensuring participants complete the entire survey. For this reason, the easier and (likely) more interesting sections were placed first, and within each section easier questions were placed first. Within each section, a qualifier question was asked first (if relevant). For instance, the first question in the Version section asked whether they keep multiple versions. If they answer no, the rest of the section becomes non-applicable and they can skip to the next section.

The question chosen to be the first question in the survey is particularly important, since it can set the tone for the entire survey. The first question was selected for its salience, ease of answering and its applicability to all survey respondents (Dillman, 2000).

Question 1: How organised would you say your documents and folders usually are?

- Very organised
- Somewhat organised
- Not very organised
- Not at all organised

In Question 1, there is deliberately no midpoint in the scale. This forces the respondents to categorise themselves as either organised or not organised rather than being neutral.

Question 2 is a block of ten attitude statements about personal document management. The respondents answer using a five point Likert scale ranging from 'strongly disagree' to 'strongly agree' with a midpoint of neutral.

These two questions on attitude are deliberately placed at the beginning of the survey in order to get an unbiased response. If they were placed at the end of the survey, the response might be coloured by having spent time thinking about issues raised by the other questions (such as failing to find files when searching, or having problems with versioning) (Dillman, 2000).

Section 2 asks the respondent to imagine Microsoft consulting them about potential changes to file management in Windows and then asks the following three questions:

What would you tell them was the best thing about the current system, the one you would definitely like to keep?

What would you tell them was the worst thing about the current system, the one you would definitely like to remove?

What would you tell them was the most useful new thing they could add to the system?

These questions were also deliberately placed at the beginning so that questions about particular features of the Windows file management system didn't influence the response. Open questions like these are more difficult to analyse and compare, because the range of options is not known in advance. In this case, the range of possible answers is unknown and potentially limitless. These questions also have the advantage of getting responses in the respondents' own words.

The demographics section was placed near the end, since many people find demographic questions boring (Bourque & Fielder, 1995). Starting the survey with the demographic questions can put people off completing the survey altogether.

The final section is a single freeform text question asking the participants if they had anything else they wished to say, allowing for them to possibly clarify any previous responses or add any relevant insights. The inclusion of a question like this is recommended in any questionnaire (Bourque & Fielder, 1995), since it allows respondents to vent their feelings about the topic or about the questionnaire.

These are the sections that were used in the final survey:

1. Attitudes (2 questions)
2. New System Features (3 questions)
3. Desktop (9 questions)
4. My Documents (2 questions)
5. Creating and Naming Files (7 questions)
6. Creating and Naming Folders (2 questions)
7. Locating Documents (4 questions)
8. Searching (5 questions)
9. Viewing Documents (3 questions)
10. Versions (4 questions)
11. Copies (2 questions)
12. Deleting and Backup (5 questions)
13. Demographics (11 questions)
14. Comments (1 question)

The wording and order questions were iteratively improved and tweaked by the researcher in consultation with colleagues until a stable set of 60 questions was established. The full set of survey questions and possible responses are provided in **Appendix G**.

5.1.1 Questionnaire Design

A self-administered online questionnaire was selected for delivery of the questions. Self-administered questionnaires have the advantages of allowing a large number of responses in a short time, with lower cost compare to other methods, the potential for wider coverage and a larger sample (Bourque & Fielder, 1995). This topic is well suited to self-administered questionnaire, since the topic is well contained, focussed on present phenomena, and everybody should be able to answer almost all

questions (Bourque & Fielder, 1995). Bourque and Fielder further recommend that self-administered questionnaires should not be used for exploratory studies. Since this is a descriptive study investigating current behaviour with a clearly defined scope, this study meets these requirements.

A web based questionnaire was chosen since this allows the participants to run the file system snapshot software, which would not be possible with a paper or telephone based survey. Web based questionnaires also have other advantages, including shorter timelines, lower costs and the opportunity to ensure fewer data entry errors (Porter, 2004). However, using the web as the delivery mode could have potential for introducing coverage error if not all members of a population have access to the web. Fortunately, in this case the caution does not apply, since all members of the target population have internet access.

There is yet no consensus in the literature on whether or not there are differences in response rates between web surveys or paper based surveys. In his meta-analysis of survey response rates, Porter (2004) notes that experimental testing of web response rates has sometime resulted in the web having higher response rates, and sometimes in the paper-based surveys having higher response rates. In most cases when both web and paper based surveys have been used, the web surveys have had lower overall response rates, however it is not clear whether this is due solely to the mode, or due to other differences such as the number and method of contacts, and whether or not incentives were provided (usually they are in paper surveys, usually not in web surveys).

Elements of Dillman's Tailored Design technique for surveys were used in constructing the invitation email, the introduction page of the web survey, and the follow up email (Dillman, 2000). Tailored Design draws on social exchange theory to frame the interaction between researcher and participant as a social exchange based on trust and reciprocity rather than an economic exchange. This is reflected in the wording used in the contacts, such as asking for help rather than demanding compliance, and explicitly stating what they will get in return (but not in monetary terms). Troutead found that phrasing the invitation as a request for help could increase response rate by as much as 5% (2004).

This effect might be stronger because of the University setting. Many of the potential participants are themselves researchers who rely on the goodwill of others to participate in their own research. This desire to reciprocate may make the social incentive for them to participate higher.

Offering an incentive after the survey poses the risk of the participants framing the interaction as a purely economic exchange rather than a social one, and thus could actually make them less likely to participate. Providing an incentive to everyone before the survey is more effective, but due to the difficulty of doing this with a web-based survey, no incentive was offered.

One of the most significant factors to affect response rate and quality has been the number and timing of contacts between the surveyor and the participant (Bourque & Fielder, 1995; Deutskens et al., 2004; Dillman, 2000). Dillman (2000) summarises over 20 years of research into his Tailored Design Method on the subject and concludes that 4-5 separate contacts provide optimum response rates. However these results were based primarily on experience with postal surveys. With the current prevalence of spam, I chose to only send two emails – an initial invitation and a thank you/reminder email.

When Dillman did use the internet for some parts of the survey, he found that mixing the mode of these contacts e.g. sending physical letters to invite people to do a web survey) did not produce such good results. Dillman found that overall, response rate to web surveys was highest when all contacts were made by email, and in particular, mixing the mode of the contacts appeared to reduce the overall response rate.

As the survey was delivered via the web, there were three contacts to each participant: two email messages, and the website that delivers the questionnaire itself.

5.1.1.1 First Contact: Invitation Email

The first contact is an introduction email and contains an invitation to participate in the survey along with a link to the survey web site.

The text of the first contact invitation email is shown in **Figure 48**.

The list of participants' names and email addresses was sourced from the University of Auckland Business School mail server. Although the Business School has an email alias that can be used to send a message to every employee at the same time, this was not used, and instead the emails were personalised and sent individually to each participant.

Personalising the salutation increases the respondents sense of their importance and makes them more likely to answer the survey than if they believe they are just one of a number of anonymous participants (Bourque & Fielder, 1995). Personalising the invitation letter (e.g. Dear Pat, or Dear Pat Lee) result in higher response rates than when using generic salutations (e.g. Dear Staff Member). Joinson and Reips (2005) found the odds of a person responding can increase by up to 40% when the 'Dear Pat' form of salutation was used instead of an impersonal salutation. Although this effect may only be present when the surveyor is of higher status than the respondent, and indicates socially desirable responding, the balance of evidence suggests that personalisation could potentially improve response rates, and at least will not decrease them.

From: Henderson, Sarah
Subject: Personal Digital Document Management Survey

Dear Jane,

I'm Sarah Henderson, a PhD student in the Department of Information Systems and Operations Management, and I am writing to ask for your help with my research into Personal Digital Document Management. In order to improve the software support for personal digital document management, I need to find out how people currently manage their documents.

I would like you to help me with this by completing a short web-based survey. The survey has between 40 and 60 questions, and takes about **10 - 15 minutes**. The final question asks you to run some software that takes a snapshot of your file system (but don't worry, it can't see the contents of your files, and it's optional).

In return, you get a chance to see **how your file system compares** against other people who participated in this research (if you take the file system snapshot at the end of the survey).

You also get a **warm glow of satisfaction**, knowing that you are contributing valuable information that will help to inform the design of more effective personal digital document management systems.

You have received this invitation email because your name appears in the University of Auckland Business School Global Address List.

To participate in this survey, you should be:

- A staff member of the University of Auckland Business School
- Accessing the survey from your primary work computer at the University
- Using a computer running Microsoft Windows to manage your documents
- Preferably take this survey using the Internet Explorer web browser

This survey is completely **anonymous**, and your participation is **voluntary**.

If you wish to participate in the survey, please go to this web page:

<http://pddm.isom.auckland.ac.nz/survey/index.php>

Thank you very much for your time and help in making this study possible. If you have any queries or technical problems or wish to know more please phone me on the number given below or reply to this email.

Yours sincerely,
Sarah Henderson

P.S. If you do not want to participate, or believe you received this email in error, please let me know by replying to this email so I can make sure I don't send you any follow-up reminders.

Sarah Henderson

PhD Candidate - Information Systems and Operations Management
University of Auckland
Office: Old Choral Hall M3
Phone: 64 (9) 373-7599 ext 86648
Email: s.henderson@auckland.ac.nz

This survey was approved by the University of Auckland Human Participants Ethics Committee on the 9th of March 2005, for a period of 3 years from the 9th of March 2005. Reference 2005/Q/009.

Figure 48: First Contact - Invitation email sent to survey participants

Personal Digital Document Management

Survey Conducted by [Sarah Henderson](#),
Department of Information Systems and Operations Management

Thanks for helping me out!



THE UNIVERSITY OF AUCKLAND
BUSINESS SCHOOL
INFORMATION SYSTEMS
AND OPERATIONS MANAGEMENT

I'm Sarah Henderson, a PhD student in the Department of Information Systems and Operations Management, and I'm studying how people manage their personal digital documents. I would like you to help me find out about document management practices by completing this survey.

What you have to do: Answer questions about how you manage and work with your documents on your computer. There are between 40 and 60 questions (depending on your answers), but they are relatively quick so the survey should only take you about 10 minutes.

The final question asks you to run some software that takes a snapshot of your file system (but don't worry, it doesn't look at the contents of your files, and its optional).

What you get in return: The chance to see how your file system compares against other people who participated in this research (if you take the file system snapshot at the end of the survey).

A warm glow of satisfaction, knowing that you are contributing valuable information that will help to inform the design of more effective personal digital document management systems.

Are you eligible to do this survey: To take this survey, you need to be:

- A staff member of the University of Auckland Business School
- Accessing the survey from your primary work computer
- Using a computer running Microsoft Windows to manage your documents
- Preferably take this survey using the Internet Explorer web browser

The following procedures are in place for your protection:

- **Anonymity.** All data obtained is completely anonymous. Information will be derived from the data collected in such a way that individuals can not be identified.
- **Voluntary Participation.** Participation in this research is voluntary and you may decline to take part without giving a reason.

Thank you very much for your time and help in making this study possible. If you have any queries or wish to know more please phone me on the number given below or write to me at the address given below.

Yours sincerely,

Sarah Henderson

[Start >>>](#)

This survey was approved by the University of Auckland Human Participants Ethics Committee on the 9th of March 2005, for a period of 3 years from the 9th of March 2005. Reference 2005/Q/009.

Figure 49: Second Contact - Introduction page on survey website

The invitation emails were sent out early in the mornings over a two day period consisting of a Tuesday and Wednesday, since the time of day and day of the week the invitation is sent may also be a factor in the response rate. Saturday and Sunday have the lowest response rates, and Tuesday

afternoon and Wednesday morning were found to have the highest rates (Faught, Whitten, & Green Jr., 2004). Trouteaud (Trouteaud, 2004) found that higher response rates were achieved when the email was sent before the start of business, than when sent at lunchtime.

The survey began during the 6th week of the semester (out of 12), two weeks prior to a one week break. The reminder emails were sent out in the 7th week, and the survey was closed at the end of the break.

5.1.1.2 Second Contact: Survey Website

The second contact is the introduction page of the website itself, which is shown in **Figure 49**. This contact continues the theme of social exchange by starting with the words ‘Thank you for helping me out’. It repeats some of the information in the initial invitation email, with more information about the voluntary nature and the anonymity of the survey.

5.1.1.3 Third Contact: Thank You Email

The third contact is a combination of thank you and reminder email (shown in **Figure 50**). Like the first email, this email is also personalised and individually sent.

Because the survey is truly anonymous, it is not actually possible to determine who has answered the survey and who hasn’t. Hence, the letter was phrased as ‘Thank you for helping me out with this survey, but if you haven’t, there’s still time’, and providing a second link to the survey. This also draws on the social exchange theory, since thanking people for something that they haven’t yet performed places a social obligation to complete the task.

Deutskens, Ruyter et al, (2004) found in their test of web survey responses that they had received over half of their responses within 3 days of sending the invitation, and the average time taken to respond was 6.6 days. They also found that follow-up emails sent after one week were as effective as emails sent after two weeks.

This email was sent a week after the initial invitation email was sent. Many people get such a large number of emails that an email over a week old is likely to have been discarded and ignored.

5.1.1.4 Survey Website Design

The questionnaire was created in HTML (hypertext mark-up language) and delivered through a web server to which all university staff had access.

The survey was a sequence of four pages: the introduction page, the survey page, the snapshot page and the thank you page.

From: Henderson, Sarah

Subject: Thank you for helping me out with my Personal Digital Document Management Survey

Dear Jane,

Thank you for very much for helping me out by completing my survey on Personal Digital Document Management.

I really appreciate you taking the time to help me out with my research, and I look forward to using all the data I collected to try and improve the design of personal digital document management systems.

Because the survey was anonymous, I can't actually tell whether or not you have already completed the survey. If you haven't completed it yet, I still need more responses, and there's still time. The survey will be available until 5pm on Friday 10th June, and can be accessed at <http://pddm.isom.auckland.ac.nz/survey/index.php>.

Yours sincerely,

Sarah Henderson

Sarah Henderson

PhD Candidate - Information Systems and Operations Management

University of Auckland

Office: Old Choral Hall M3

Phone: 64 (9) 373-7599 ext 86648

Email: s.henderson@auckland.ac.nz

Figure 50: Third Contact - Combination of thank you and reminder email sent to participants one week after the initial contact.

The pages were designed to be only 640 pixels wide to ensure that even participants with the smallest screen resolution will be able to view the survey. The restricted width (compared with a full 800x600 or 1024x768 or bigger display) makes the questions more compact and easier to read on a screen. Response choices for most questions were usually placed vertically, and the controls used were radio buttons, textboxes and buttons that would be easily recognisable by any Windows XP user.

Introduction Page

The introduction page explains to the participant what is involved in the survey, and reassures them that participation is voluntary, and that the survey is anonymous (as per University of Auckland ethics guidelines). The time taken to complete the survey is clearly stated. It is important that this is accurate, otherwise respondents may abandon the survey if they feel they were misled about how long it would take (Bourque & Fielder, 1995), and therefore it was determined after initial pre-testing by the researcher and several colleagues.

The introduction page also includes information on how to contact the researcher if they have any questions or concerns. At the bottom of the information is a clearly marked Start button for the participants to press to begin the survey.

The wording of the introduction was chosen to take advantage of social exchange theory to increase participation. The heading text is 'Thanks for helping me out!' and the section on 'What you get in return' talks about a 'warm glow of satisfaction' which attempt to position the act of filling out a survey as doing a favour.

Survey Page

All 60 survey questions are on the same page. The page has the same header and footer and overall look and feel as the previous page.

Survey Progress
 Answered: 0
 Remaining: 60

Personal Digital Document Management

Survey Conducted by [Sarah Henderson](#),
 Department of Information Systems and Operations Management

Attitudes

For all the questions in this survey, please think about the documents and folders stored on the hard drive of your primary work computer.

1 How organised would you say your documents and folders usually are?

☐ Very organised
☐ Somewhat organised
☐ Not very organised
☐ Not at all organised

2 How much do you agree or disagree with each of the following statements?

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I feel my documents are well organised to suit my working habits	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I sometimes get annoyed at the time taken to locate my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that someone else would be easily able to find things in my system	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I sometimes wish there was a better way to organise my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am quite happy with the way I manage my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would rather spend less time organising my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The time I spend organising my documents is worth it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would be embarrassed to show someone how my documents are organised	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think it is important to have my documents well organised	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If someone showed me a better way to organise my documents, I would probably change the way I do it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 51: Survey page showing first two survey questions

Each section has a header and brief description/directive explaining the context of the questions. Each question is clearly numbered in the left margin. The font face, colour and style of the text have been chosen to maximise legibility, and the survey has been coded in a way that allows users to increase the text size using their browser's text-resizing facility if necessary.

Sets of radio buttons were used for most of the answer choices, since these are least affected by primacy (Couper, Tourangeau, Conrad, & Crawford, 2004). In two cases, drop down boxes were used, but only in situations where the number of options to choose was large, and the participants were not making a value judgement but just indicating a state: choosing department and job title.

As the survey respondents are all staff members of a Business School, and are all supplied with a computer for their work, they are all technologically literate enough not to require instructions on things such as how to work a combo-box, how to use radio buttons. Providing this type of instruction would be more likely to insult most users than assist them.

For questions that have multiple statements with Likert scale responses, these have been grouped into a table, with each alternating row shaded to help visually match up the statement with the correct line of responses.

Some questions on the survey are not applicable to all respondents. For instance, if a respondent doesn't use the Desktop, then none of the questions about the use of the Desktop are applicable. It is generally recommended that the use of skips should be minimised in self administered questionnaires (Bourque & Fielder, 1995). However, they are better than forcing respondents to answer questions which are not applicable.

My Documents

15 Do you use the system created My Documents folder to store your documents?

☒ Yes [SKIP to Question 17](#)

☐ No

16 (If No) Why not?

☐ Must be stored elsewhere to be backed up

☐ Must be stored elsewhere so they will not get backed up

☐ Old habit of storing elsewhere

☐ I don't know

☐ Other:

Creating and Naming Documents

17 How do you usually create new documents?

☐ I open an existing document and use Save As

☐ I copy and rename an existing document in the file system

☐ I right-click inside a folder and choose New

☐ I use the Quick Start or Office Toolbar

☐ I use the New Office Document from the Start Menu

☐ I open the application

Figure 52: Survey question showing skip link and disabled question depending on user response.

The web based delivery of the survey makes presentation of this situation easier, since the non-applicable options are automatically disabled, and the respondent is given a link to skip to the next relevant question. An example of this is shown in **Figure 52**, where the respondent has answered yes to Question 15. When they select 'Yes', the skip link appears, and the question below is disabled so that the user cannot respond. If the respondent changes their response to no, the skip link will disappear and the options for question 16 will be re-enabled.

In the left hand side of the page is a small progress indicator window. This progress indicator moves down as the user scrolls down the page, so is always visible. This shows how many questions have been answered and how many are remaining and shows this progress graphically with a progress bar.

Figure 53 shows the indicator at three different stages during the survey. At the start, it shows that no questions have been answered, and there are 60 possible questions remaining. As the respondent answers each question, the progress is updated immediately to reflect the current progress. If the respondent chooses options that result in some questions becoming inapplicable, those questions are removed from the count of remaining questions. At the end of the survey the progress bar is full and a count of all applicable questions answered is displayed.

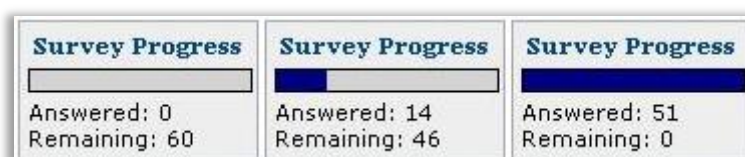


Figure 53: Survey progress indicators at three different times during the survey, (a) at the start, (b) partway through and (c) at the end.

Once the respondent has reached the end of the survey, there is a prominent button marked 'Continue' which takes them to the next page in the survey. No attempt was made to force the users to answer every question in the survey. While this was technologically feasible, it is possible that some users might object to answering certain questions and might abandon the survey rather than go back and answer the question. The trade-off is that not all participants will necessarily answer every single question, and thus there will be some missing data in the survey.

Snapshot Page

The snapshot page (shown in **Figure 54**) explains to the respondent what the information the file system snapshot will record on their computer and then provides four step instructions for completing the snapshot (although it is designed to be as easy-to-use as possible).

When the respondent clicks on the link to open the snapshot, they are prompted by their browser whether they want to run the program. If they choose to run it, the File System Snapshot application (shown in **Figure 55** below) will run.

Personal Digital Document Management

Survey Conducted by [Sarah Henderson](#),
Department of Information Systems and Operations Management

File System Snapshot

That's all the questions completed. Now, you get to see how your file system compares against everyone else's. To do this, my software will examine your file system and look at the following things:

1. The structure of the file system hierarchy
2. Folder name and date created, last accessed and last modified
3. File name, size, date created, last accessed and last modified
4. Total number of files, folders, and the depth of your folder structure

Don't worry, it doesn't look at the contents of any file, and does not collect any information that can be used to identify you. It will probably take between 1 and 5 minutes to run depending on the number of files and folders you have. As it runs, it will display some running statistics about your file system.

Step One Open the File System Snapshot program by clicking here: [Open](#)
When prompted, choose to Open or Run the program. If necessary, it will first install the .NET framework that it requires to run, and then will start the snapshot program.

Step Two If you store your documents somewhere other than the Desktop or the My Documents folder, add that location to the list. (Don't choose the whole C:\ drive, or you'll be waiting all day - only include document folders).

Step Three Press the green "Start Snapshot" button, and watch the statistics change.

Step Four When the snapshot says "Snapshot Complete", close the window, then press Finish below to finish the survey.

Finish > > >

This survey was approved by the University of Auckland Human Participants Ethics Committee on the 9th of March 2005, for a period of 3 years from the 9th of March 2005. Reference 2005/Q/009.
Please contact me at s.henderson@auckland.ac.nz if you have any questions about this research.

Figure 54: File System Snapshot page of the questionnaire.

By default, the File System Snapshot software includes the Desktop and My Documents folders. The participant is instructed to select the folders where they keep their documents, and there are clearly labelled Add and Remove buttons for them to make adjustments. The very prominent green Start Snapshot button begins the process of taking the snapshot. The comparison data is a live summary of all other survey responses so far. At the commencement of the survey period, the comparison values are the averages from the interview snapshots. The metrics and comparisons update live as the program is running, to keep the respondent interested and make them less likely to cancel the snapshot.

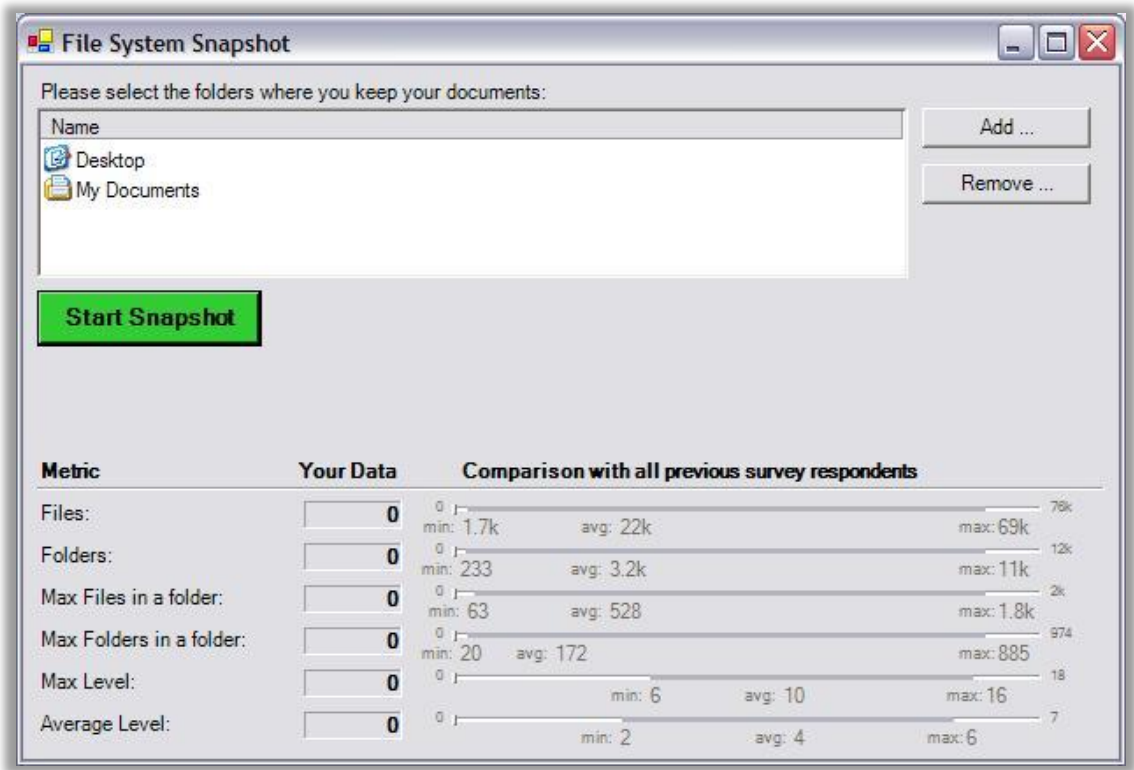


Figure 55: File System Snapshot software waiting to run.

Figure 56 shows the software with metrics and comparisons displayed, and indicating that the snapshot is complete. The snapshot software records all the data to a database on the same server that hosts the survey website.

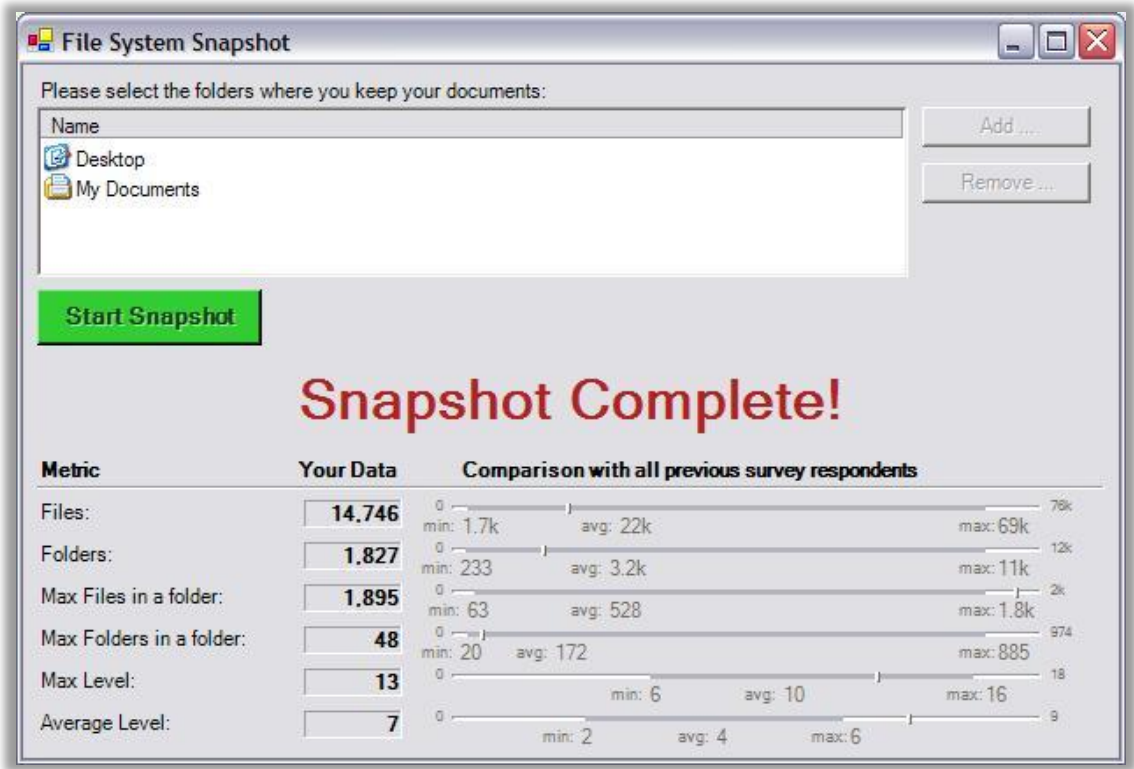


Figure 56: File System Snapshot software after running.

Thank You Page

The final page of the survey announces that it is finished and thanks the respondent for their participation (as shown in **Figure 57**). It also gives contact details for the respondents to contact the researcher if they have any queries.



Figure 57: Questionnaire Thank You page

5.1.2 Testing

As this questionnaire has never been used before, pre-testing was essential (Boudreau et al., 2001). Pretesting involves reviewing the questions and providing feedback on whether the questions are easy to understand, whether they had any difficulty with any aspect of the questionnaire, and whether they thought anything was missing (Bourque & Fielder, 1995). The questions were pre-tested with four colleagues. They were specifically asked to comment on any ambiguities in the question and response wording. As a result of the pre-testing, some changes were made to the wording of some questions.

Following the pretesting, a pilot test of the survey website was conducted by three colleagues. In addition to testing the survey instrument, a pilot study includes testing all the delivery and data collection procedures, as well as being used to produce an estimate of how long it would take the complete the survey (Punch, 2003). This included a test of sending the invitation emails, the website itself (including the skip logic and progress indicator) and the file system snapshot software.

5.1.2.1 Reliability and Validity

It is important for any measurement to be reliable and valid. Reliability expresses whether or not the research gives consistent and reproducible results over time, whereas validity expresses how well an instrument measures what it sets out to measure (Litwin, 1995).

There are four general techniques for estimating the reliability of a measure (Trochim, 2002):

Inter-observer reliability assesses the extent to which different people agree on a phenomenon. It is not applicable here because the individual is the 'observer' of their document management practices.

Parallel-forms reliability ensures that two or more forms (e.g. different question phrasings) of the same measure are consistent with each other. This is also not applicable, since it is not possible to add alternate forms of the questions to the survey without making it long and therefore making it less likely to be completed.

Internal consistency reliability ensures that groups of items thought to measure different aspects of the same concept are in fact related to each other. It is measured by calculating Cronbach's alpha (Litwin, 1995). As the questions were designed to cover a broad spectrum of document management behaviour, there is little scope for assessing this. However, on the question of attitudes to document management, multiple statements expressing the same idea have been constructed and so the Cronbach's alpha will be calculated for those in order to check the reliability.

Test-retest reliability ensures that a measure is consistent over time. It is measured by having the same respondents complete the survey at two different points in time and correlation coefficients are used to compare the two sets (Kitchenham & Pfleeger, 2002). It is important that the questions are not asking about things that can change rapidly over time, and that having recall of the previous test would not change the answers.

In line with good practice, the two pilot study participants also took the survey approximately six weeks after first taking the survey during the pilot test. Their free text responses were not considered (since they had not entered any during the pilot testing), but their other responses were compared to see how different they were. For all nominal questions (asking about use of specific features), their answers did not change from the pilot to the actual survey. For the other questions, while there were some differences in assessment of frequencies and attitudes, the correlation between the test and retest was 0.94 for the first tester and 0.97 for the second tester.

Internal validity expresses how well an instrument measures what it sets out to measure (Litwin, 1995). There are a number of different forms of validity (Trochim, 2002), some of which are applicable to this research and some are not:

Face validity is the least scientific form of validity and involves a cursory review to see whether the items appear to measure what they set out to measure. The survey certainly has face validity, based on the feedback of other researchers during pretesting phase.

Content validity is similar to face validity except that the reviewers have knowledge of the subject matter and it takes a more systematic approach. Content validity is ensured through the process of

developing the survey questions directly from the personal document management conceptual framework produced from the interviews. This ensures that each concept was reasonably well covered (as much as possible within question limits).

Concurrent validity measures how well an instrument compares to some other instrument that is acknowledged as valid. Since no other instrument exists, this is not possible. However, the file system snapshot provides a partial assurance of concurrent validity, since participants responses can be checked against their document structure.

Predictive validity is the ability of the instrument to predict future outcomes, and isn't relevant to this study, since this is a cross-sectional study rather than longitudinal.

Convergent validity implies that an instrument is correlated with other instruments measuring the same thing, and is not applicable due to lack of other instruments.

Divergent validity means that an instrument should not correlate too closely with similar but distinct instruments, which again is not applicable due to lack of other instruments.

External validity is the validity of being able to generalise results to a wider population. It can be assured only if the sample was representative of the population to which the findings are being generalised (Fink, 1995b). In this case, no attempt is being made to generalise to the population of all knowledge workers, but instead, the results are being used to validate the conceptual model developed during the study.

5.1.2.2 Minimising Error

In any research, there are possible sources of error that must be controlled (Dillman & Bowker, 2001). There are four main sources of error that are possible with sample surveys:

Coverage error occurs when there is not a known non-zero probability of each participant being included in the sample drawn to represent the population, meaning that sampling frame does not match the population to which the results are to be generalised. In this case, the results are not being generalised to a wider population so this is not an issue.

Sampling error arises due to the measurements being made on only a proportion of the population and not the whole population. Because The University of Auckland was chosen as a convenience sample and not a random sample, there is no reason to think that results generated here can necessarily be generalised to other universities, let alone to all knowledge workers. Again, in this case, the results are not being generalised to a wider population.

Non-response error is the error that occurs when the people who do not respond differ in some systematic way from those who do respond. Non-response in the web environment can be caused by

technological problems, such as those with different software, less powerful hardware or slower connections being unable to access the survey at all, by respondents giving up because they didn't know how close to the end they were, and being forced by the technology to answer every question, even when not applicable or appropriate (Dillman & Bowker, 2001). In this survey, many of these concerns are dispelled due to the technology environment. All the staff computers in the business school are less than three years old (kept that way by 3-year leases), and have the same hardware platform. All are initially preinstalled with the same software, although staff have the ability to alter the configuration if they desire. Additionally, extreme care is taken to use only standard web technologies for most of the survey, and where that is not possible (in administering the file system snapshot), alternative methods of running the snapshot have been provided.

Measurement error occurs as a result of inaccurate responses that arise from badly worded questions and question options, or poor survey administration. One notable source of this on the web is due to the variety of software and hardware configurations possible, especially different web browsers and screen resolutions. It requires care to be taken to ensure that the survey format does not alter under different conditions in such a way that the measurement scales are compromised. In this survey, only standard, well-supported web technologies were used, and the survey was tested out in all recent versions of the major browsers to ensure that the measurement scales weren't compromised.

5.2 SURVEY RESULTS

At the time of the survey, the business school staff directory contained 528 entries. The initial invitation email was sent to these people. 7 emails were returned as undeliverable. These were checked to make sure there were no spelling errors, but it seems that those staff had ceased employment, and their email accounts had been deleted prior to their removal from the staff directory. 6 people responded that they were not staff members although their email account was still active. Of those 6, four indicated they had ceased employment, and 2 indicated that they had email accounts for other reasons. 25 auto reply emails were received indicating that the person was away and would be away for at least the following week.

This leaves a total of 490 staff members who were able to participate. 7 people replied to my email declining to take part. One did not give a reason, but 5 said they were too busy, and 1 said he used an Apple Mac rather than a Windows PC. 116 responses were received, although one of these was a duplicate response and was eliminated from the analysis, making 115 valid responses. This gives a response rate of 23.5%. Of these 115 survey responses, 78 also completed the file system snapshot (a response rate for that part of 15.9%). Response rates for mail surveys are frequently in the range of 20-40% (Punch, 2003).

5.2.1 Questionnaire Section 1: Attitudes

The first section was designed to gauge the participant's attitudes and satisfaction with their personal document management practices.

85% of respondents report that they feel their documents are either very organised or somewhat organised, and only 1 respondent considered themselves not at all organised.

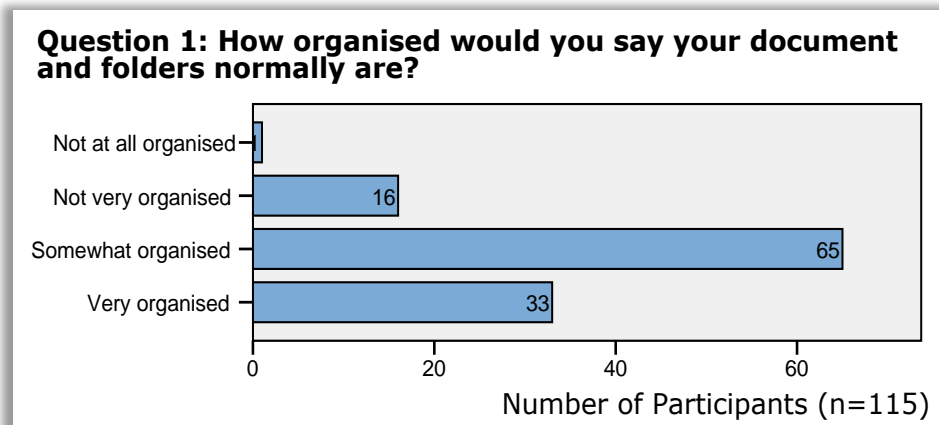


Figure 58: Question 1 - Bar graph showing how organised participants describe themselves as

The items in question two were intended to capture how the participants feel about their personal document management practices. Because these are anticipated to be aspects of one or more underlying concepts, rather than analyse these individually, I will use a principal components analysis to create a single score. The respondents answered all statement in question 2 using a 5 point Likert scale with the items Strongly disagree, Disagree, Neutral, Agree, Strongly Agree. **Table 11** below lists the statements the respondents were asked to consider.

Table 11: Attitude statements on Survey Question 2

Item	Statement
a	I feel my documents are well organised to suit my working habits
b	I sometimes get annoyed at the time taken to locate my documents
c	I think that someone else would be easily able to find things in my system
d	I sometimes wish there was a better way to organise my documents
e	I am quite happy with the way I manage my documents
f	I would rather spend less time organising my documents
g	The time I spend organising my documents is worth it
h	I would be embarrassed to show someone how my documents are organised
i	I think it is important to have my documents well organised
j	If someone showed me a better way to organise my documents, I would probably change the way I do it.

Principal components analysis requires a high degree of inter-correlation between the variables. Of the 45 pairs of items, 33 were significantly correlated (at least to the 0.05 level of significance). Bartlett's Test of Sphericity is highly significant, confirming the presence of inter-correlations. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy has a value of 0.753, again confirming the inter-correlation criterion is met.

An initial un-rotated factor analysis produced three factors with eigenvalues greater than 1, however no variables loaded most heavily on the third factor, and the scree test indicated a two-factor solution (Ho, 2006). Therefore, a two factor solution was used. An oblique rotation was applied to this two-factor solution, since the goal is to discover theoretically meaningful factors and it is expected that there will be correlations between factors (Hair et al., 1995, p. 110). The following table shows the pattern matrix produced using an oblimin rotation on a 2 factor solution:

Table 12: Factor analysis pattern matrix (Oblimin rotation)

	Component	
	1	2
a. I feel my documents are well organised to suit my working habits	.602	
b. I sometimes get annoyed at the time taken to locate my documents	-.679	
c. I think that someone else would be easily able to find things in my system		.508
d. I sometimes wish there was a better way to organise my documents	-.729	
e. I am quite happy with the way I manage my documents	.769	
f. I would rather spend less time organising my documents	-.718	
h. I would be embarrassed to show someone how my documents are organised	-.697	
i. I think it is important to have my documents well organised		.786
j. If someone showed me a better way to organise my documents, I would probably change the way I do it		.765

Extraction Method: Principal Component Analysis.

Rotation Method: Oblimin with Kaiser Normalization (converged in 9 iterations)

Only loadings greater than 0.5 are shown

Total Variance Explained: 53.3% (Component 1: 39.94%; Component 2: 16.35%)

Item g is not shown as the loadings on both factors are less than 0.5. The items that are phrased with negative statements about document management load negatively on the first factor, while the other statements load positively. All are measures of how dissatisfied or satisfied users are with their document management practices, so this factor will be called Satisfaction.

The last two items do differ conceptually from the others, in that they tend to reflect concern about appearances and therefore perhaps indicate external pressure for document management rather than internal satisfaction. Thus component 2 seems to reflect external pressures surrounding document management rather than the user's personal satisfaction, and this factor will be called External Influence.

A factor score created from factor 1, Satisfaction, will be used as a dependent variable to indicate how satisfied and happy participants are with their document management practices. It will be reversed so that positive values indicate higher than average levels of satisfaction and negative values indicate lower satisfaction. The standardised value of this metric ranges from -2.77 to 2.57. A value of 0 signifies an average level of satisfaction. A Q-Q plot indicates that the metric is normally distributed, and thus can be used for parametric analysis.



Figure 59: Normal Q-Q Plot of Satisfaction Factor Analysis Score

Satisfaction is significantly correlated with Question 1, the participant's self-report of how organised they believe they are ($r = 0.67$). Participants who rate themselves as more organised are happier with their document management practices.

5.2.2 Survey Section 2: New System Features

This section first asked the participants to "imagine Microsoft are changing the way Windows allows you to manage your documents and folders (including both the file system and the Desktop), and they have asked you for your opinion." With this prompting, they were then asked three questions asking them to identify the best feature, worst feature, and most desired new feature for the system.

5.2.2.1 Question 3: Best Feature

Question 3 asked: "What would you tell them was the best thing about the current system, the thing you would definitely want to keep?" Fifteen participants did not respond to this question, and a further two answered that they weren't sure.

Some people didn't appear to like anything about the current system. Two participants said there was "*nothing*" good about it. It is not clear whether this means they think there is nothing good at all about the file system, or if they just cannot identify one particular feature as being better than the others. Others were clearer, with one saying "*I'd struggle to find one good thing*", and another commented that "*nothing is very intuitive in Microsoft*". One person simply stated "*I hate the Desktop*".

On the other hand, several other participants had unsolicited praise for the system, with comments like "*The system appears to be totally rational to me*" and "*that it's quite intuitive*", and "*I am happy the way things are*".

The most common response theme was that the hierarchy or folder tree was the best feature. Twenty four participants listed this as the thing they would most like to see kept in a new file system design. The terms participants used to convey this varied. Typical examples are "*hierarchy*", "*hierarchical structure*", "*tree view*", "*folder tree*", "*having folders and subfolders*" and "*the folder system*". Seven participants explicitly mentioned Windows Explorer in answering this question.

A closely related theme was the flexibility, control and customisability provided by the hierarchical system of folders and subfolders. Some typical comments were "*I like the choice that I can establish folders and subfolders any time*", "*the folder system give you lots of flexibility and allows you to edit you[sic] filing system easily*," "*creating and managing my own folder structure*" and "*having easy access to seting[sic] up folders*".

Also related to this were two participants who identified the drop down file system menus which allow the tree to be traversed. These are found in some file open and file save dialog boxes, for example, "*Folders and sub-folders path system accesible[sic] from pop-down menus*"

Five participants mentioned long file names as a best feature. These have been available in Windows Operating Systems since Windows 95. Prior to this, Windows and DOS file names had been restricted to 8 characters with a 3 character extension. The participants who rated this as a feature presumably had used those previous operating systems. Users with less computing experience probably wouldn't think of long file names as being a particularly notable feature because most of their computer life has been spent in systems where long file names are the norm. These five participants had an average of 22.2 years of computer experience (compared to an average of 14.3 years for the whole sample). Additionally, 4 out of the 5 are over 50 years old (the other being in his 20s).

Two people cited file extensions as the best feature. One answering that the best feature was, "*the easy way for files type identification*". Another said that the best thing was linking file types with application programs, which is done via file extensions. File extensions also determine the icons that are used to display a document, and three people cited icons as the best feature.

Eight participants identified the ability to search as the best feature in the current system. Three people mentioned sorting as their favourite feature. One specifically referred to alphabetic sort, and another identified date and name sorting as being useful.

Three participants mentioned the Desktop, with one saying *“I like to utilise the desktop and find it works well for me. I organise and manage my files from it”*. Another simply said *“keep the desktop as it is”*. Two people also said the My Documents folder was the best feature.

Two people identified shortcuts as the best feature, one liked thumbnails and one liked the ability to easily drag and drop between folders.

One participant said familiarity was the only good thing about the system. Another didn't give anything specific, but explained that they don't want to have to learn a new approach because of the time investment involved.

5.2.2.2 Question 4: Worst Feature

This questions asked participants what they would tell Microsoft was the worst thing about the current system, the thing they would definitely like to remove or change? Twenty-two participants didn't respond to this question. Of the 93 who responded, thirteen indicated that there was nothing they could identify, or that they couldn't think of anything, and a further seven said they didn't know.

Eleven people mentioned the search as being the worst feature, with typical comments being *“text search is quite useless”*, *“search function is too slow”*, *“searching is slow and tedious”*. Two comments indicated that search was better in MS-DOS or previous versions of Windows, and one person mentioned they had recently started using Google Desktop out of frustration with the lack of a user friendly search function.

One problem mentioned was the number of clicks required to drill down into the folder hierarchy, with eight people saying things like *“difficult to access folders without clicking through lots of steps”* and *“the branching folder system is kind of tedious - would be good to somehow shortcut through,”* *“having to open mutiple[sic] folders to get to a document in a sub sub folder.”* Only one of these people thought the folder hierarchy was the best feature in the previous question, indicating that this perceived difficulty in drilling down might be a barrier to using the hierarchy.

The 255-character limit on filenames is too short for four participants, with a further participant identifying the restriction on having certain characters in file names as being a problem. Two people identified document names as lacking meaning or being insufficiently descriptive as being the biggest weakness, with a further four identifying file extensions as the worst feature. One of them further explained that the issue was the static rather than dynamic association of file types to their applications, while another said that hiding the extensions was the worst feature.

Five people thought the My Documents folder was the worst feature, with one explaining that it is because some applications have trouble recognising paths inside the My Documents folder because the path is too long. One person dislikes the fact that it is the default save location, and another objecting to all pseudo-folders, with one simply commenting *“remove My Documents, useless.”*

Three people commented that they found Windows Explorer to be *“less than user-friendly”* or *“not intuitive”*. Two others noted the lack of undo, particularly when overwriting a file. The help system was judged to be quite unhelpful by two people, with one adding *“don’t remove it though. Upgrade it.”*

Two people identified the current views offered by Windows XP to be insufficient, one saying *“I want the flexibility to look at things chronologically and by themes, and sometimes do that concurrently.”* Two people disliked the taskbar, one calling it *“awful”* and the other simply saying to remove it.

Shortcuts were the worst feature for two people although for different reasons. One said that there are too many shortcuts in too many locations and that creates confusion. The other said *“do it right, get multiple linking and symbolic linking as used in Unix.”*

One person’s objection seemed to stem not from something that hindered them, but from the way other people use the system. Their worst feature is the ability to save on the Desktop, saying it *“should be removed as people tend to just save on desktop.”*

Other features mentioned by one participant included the lack of overview, too many different ways of doing things, only having a single Desktop, having folder views without the tree for context, views defaulting to icon view.

One person simply said *“everything”*.

5.2.2.3 Question 5: Missing feature

This questions asked participants what they would tell Microsoft was the most useful new thing they could add to the system. Eighty-three participants provided a response, although eleven of those said they didn’t know and a further three said there was nothing to add.

Several respondents suggested things that it is already possible to do. For instance, one suggested *“access to the folder button on the menu bar in a similar way to the programme short-cut icons.”* It is currently possible to place any kind of shortcut on the Quick Launch toolbar or in the Start Menu, including shortcuts to folders. Another suggested a *“recently used files folder”*, which is already available on the Start Menu under the name ‘My Recent Documents’. Others who wanted search improvement specified that they wanted the ability to search within files, something that the Windows

XP Search Tool can already do for most file types, although only in the 'Advanced' search interface. The standard search interface is intended to be more usable and simpler and doesn't give this option.

Twenty respondents mentioned better search capabilities as the feature they would most like to see added/improved. Most didn't specify exactly how they wanted search to be better, although two said they wanted it more like Google Desktop, some mentioned the ability to full text search, one mentioned faster search, one mentioned being able to search documents and email together and one suggested the ability to search based on when a file was last used, rather than last modified.

Of those twenty respondents, three had nominated search as the best feature currently in Windows XP. Eight had identified the ability to create folders and a hierarchy as the best feature and the remaining nine had all identified various other features. Seven participants identified search as the worst feature currently in Windows XP. One participant identified search as the best feature, the deterioration of the search function since MS-DOS days as the worst feature and suggested better search as the most desired new feature.

Four participants referred to improved ability to create shortcuts or have documents in multiple locations, effectively creating a network rather than a tree structure. Three suggested a preview pane that allowed previews of documents in addition to pictures and video.

Several participants suggested improved sorting techniques, with one specifying that they wanted a user-controlled sort order rather than alphabetical, saying *"I constantly[sic] have to rename folders to get the order I want."*

Several participants suggested new features that require more intelligence on the part of the software. One suggested there should be a way of *"teaching the operating system what goes where so you don't have to put them there yourself"*. Another participant asked for a system that will *"automatically save your documents to a proper folder according to its content"*, and has the ability to create new folders if needed. Another wanted a system that would make them tidy their files and not allow them to get so *"out of control"*.

Some of the new feature suggestions offered by a single participant included an overview of the whole file system, the ability to easily compare files, a drafts folder for unfinished items, colour codes for different folders, arbitrary collection combining documents, emails and calendar items, the ability to group folders into *"filing cabinets"*, the ability to mark documents as being related and the ability to add metadata like tags to a file.

5.2.3 Survey Section 3: Desktop

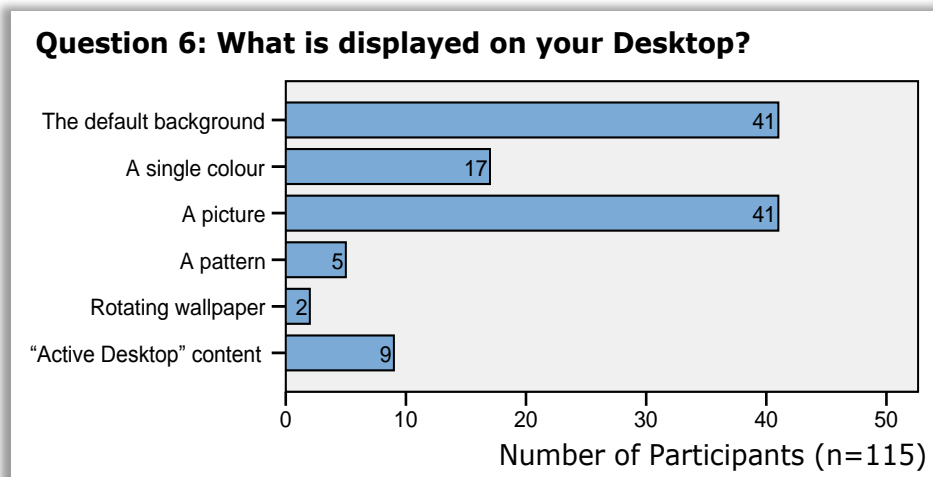


Figure 60: Question 6 - Bar chart showing what participants use as a Desktop background

36% of people haven't customised their desktop at all. Of those that have, putting a picture as background is the most common customisation, used by 36% of respondents.

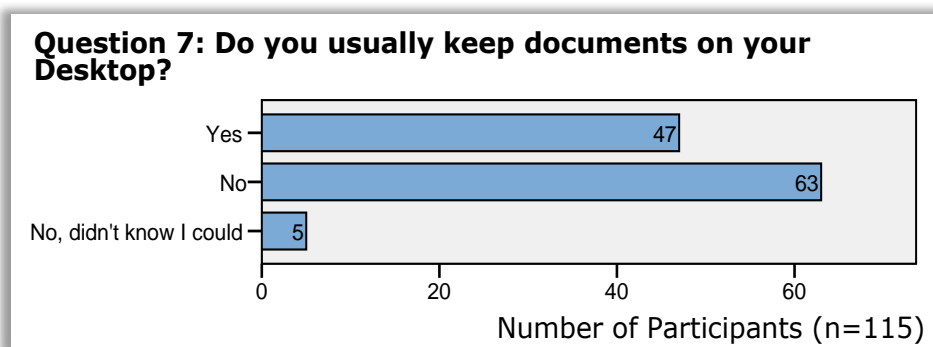


Figure 61: Question 7 - Bar chart showing whether participants keep documents on their Desktop

4% of respondents report that they didn't know they could keep documents on their Desktops. Of the remainder, 41% do keep documents on their desktop. The following questions about storage of documents on the desktop were only answered by those 47 respondents.

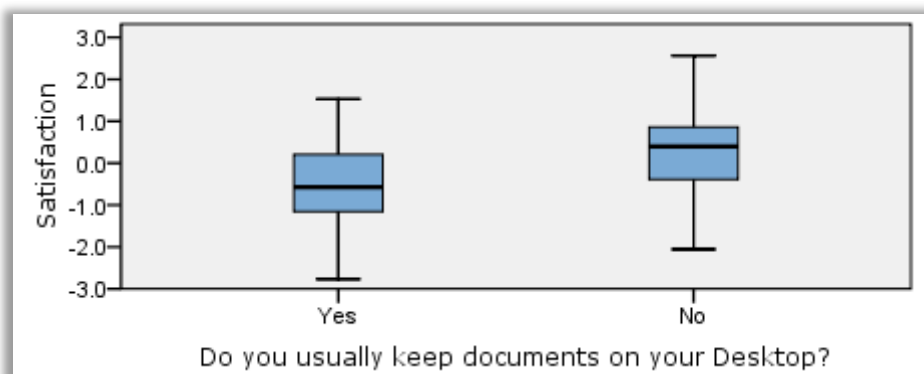


Figure 62: Box plot showing satisfaction is lower if documents are kept on the Desktop

An analysis of variance (ANOVA) indicated that there is a significant difference in satisfaction between the participants who do not keep documents on their Desktops and those who do ($f=19.52$,

sig=0.00). Satisfaction is higher for those participants who do not report keeping documents on their Desktop.

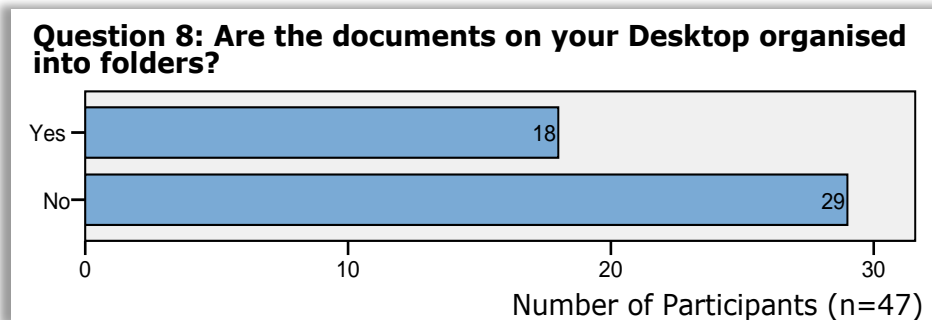


Figure 63: Question 8 - Bar chart showing whether Desktop documents are organised into folders

Of those respondents who keep documents on their desktop, 38% report that they organise them into folders while the remainder have individual documents on the desktop itself.

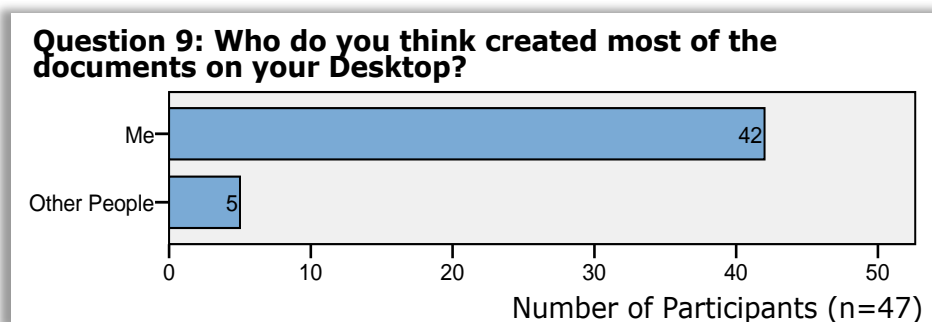


Figure 64: Question 9 - Bar chart showing who created the documents on the Desktop

89% of respondents believe that most of the documents on the Desktop were created by them, as opposed to documents sent to them or documents downloaded from the web.

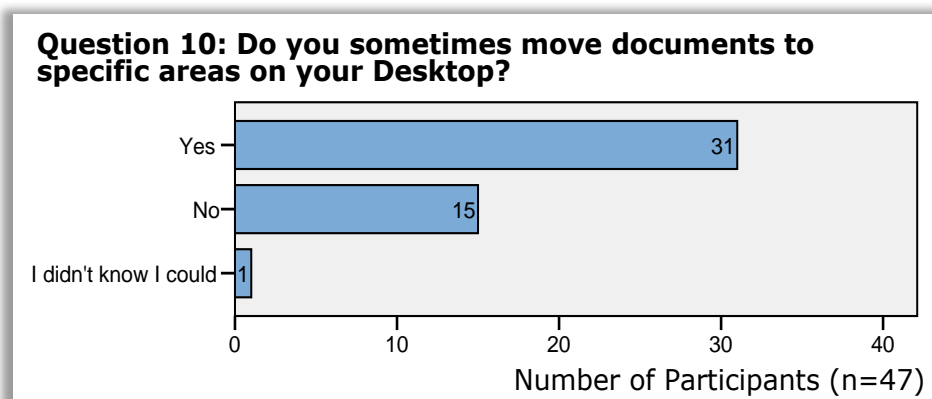


Figure 65: Question 10 - Bar chart showing whether participants use the spatial abilities of the Desktop

66% of respondents report making use of the spatial organisational ability provided by the Desktop, while only one respondent didn't know they could do this.

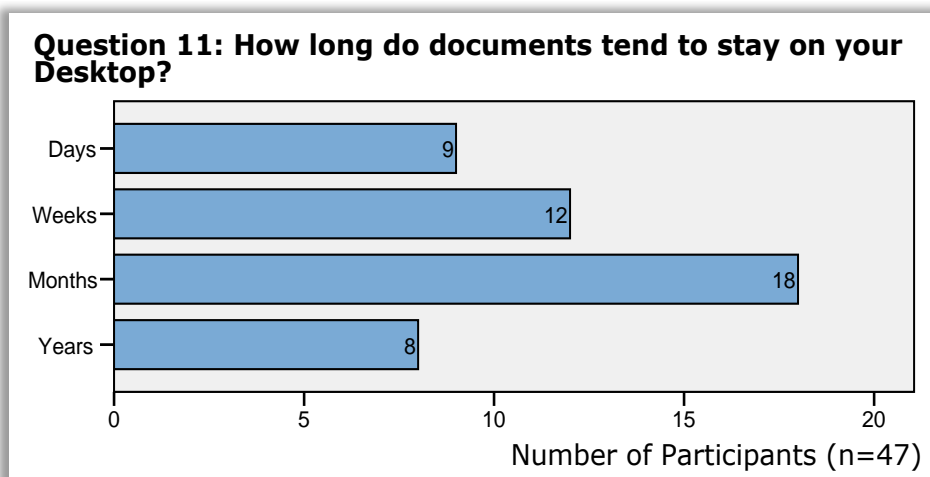


Figure 66: Question 11 - Bar chart showing how long documents tend to stay on the Desktop

There was considerable variation in the length of time people keep documents on their Desktop. The most common average duration was on the order of months, reported by 38% of respondents.

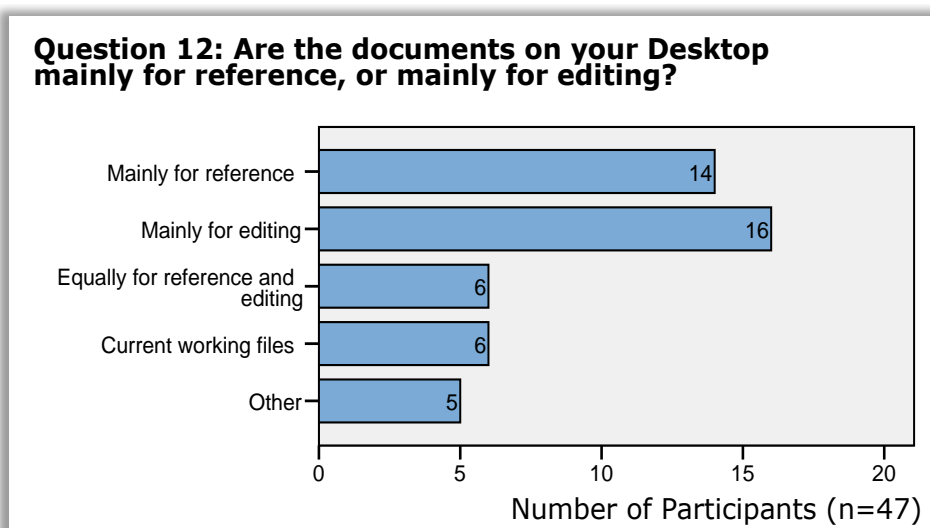


Figure 67: Question 12 - Bar chart showing the main purpose of documents on the Desktop

This question provided three options: 'Mainly for reference', 'Mainly for editing' or 'Other'. 17 respondents (36%) chose 'Other' and provided a description in the free text field provided. 6 of them indicated that they felt their documents were an equal mix of reference and editing, and another 6 indicated that the documents they keep are current working files. A further 5 people gave different free text answers, including "supposed to be editing and then they remain forever", "as a to do list", "matters needing my current attention", "mainly because they are 'different' in some way", and one person stating that all they kept were shortcuts.

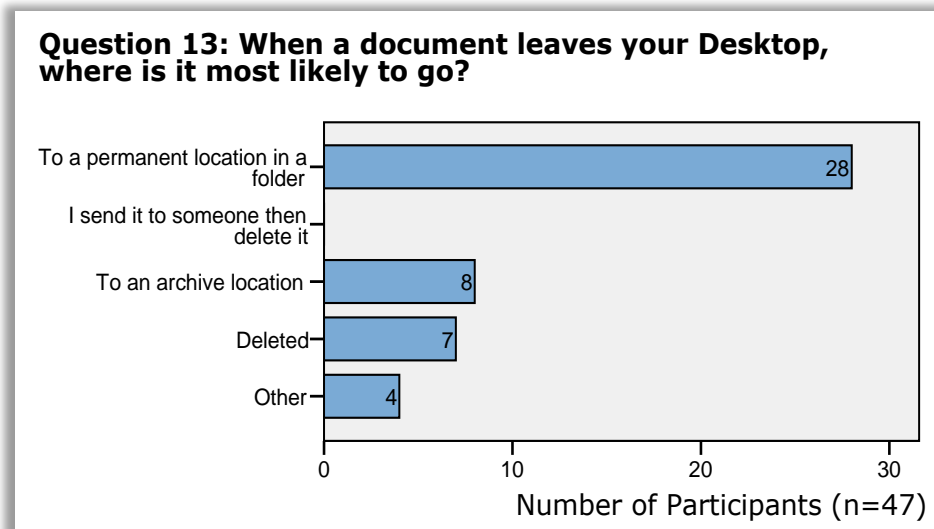


Figure 68: Question 13 - Bar chart showing where a document goes after it leaves the Desktop

Most respondents (60%) tend to move documents off the Desktop and put them into a folder structure. Of the 'Other' responses, one said they used a combination of the above, but none in particular, and another said they sometimes delete and sometimes move to a folder. Another respondent specified that they move it to My Documents for later sorting, and one said that they go into the 'Unused Desktop Shortcuts' folder that Windows XP's Desktop Cleanup Wizard automatically creates.

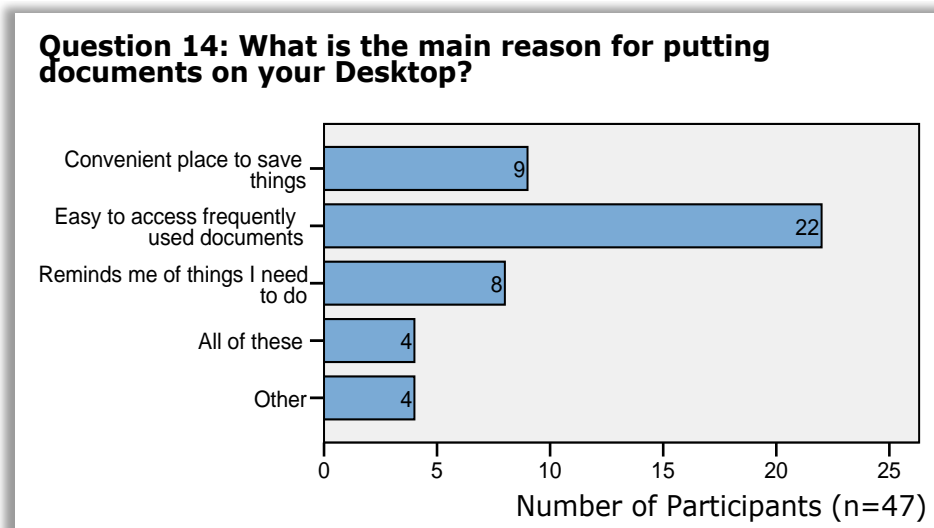


Figure 69: Question 14 - Bar chart showing the main reason for putting documents on the Desktop

'All of these' was not originally an option, but 4 respondents entered this after choosing 'Other'. Most respondents (47%) cite ease of access to frequently used documents as the main reason for using the Desktop. In the 'Other' responses, one person mentioned that they use this as their primary document storage location, another says it provides "*a focus for activities I need to do*". One mentioned they don't know where else to put it, saying it was temporary storage for them, and another saying they used it because of the ease of putting files into Cecil from the Desktop.

5.2.4 Survey Section 4: My Documents

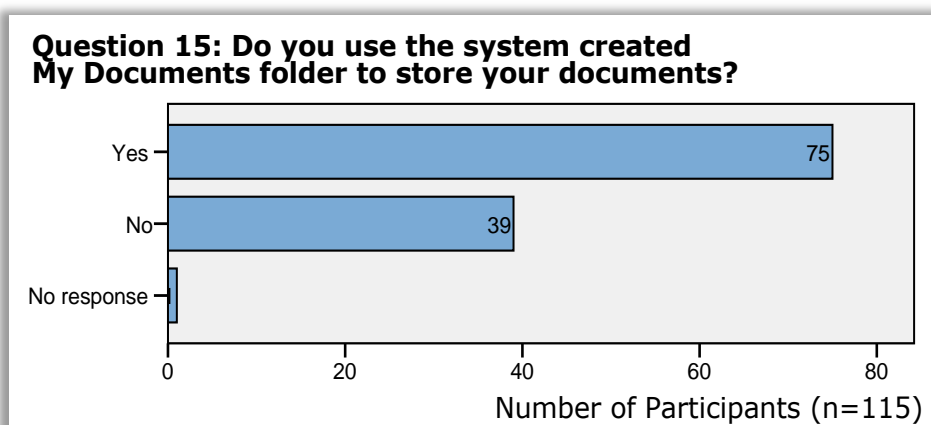


Figure 70: Question 15 - Bar chart showing whether the participants use the My Documents folder to store documents

Most respondents (65%) store documents in the system created My Documents folder. One respondent did not answer this question. The reasons why people do not use this location are explored in the following question. There is no significant difference in satisfaction between people who use the system-created My Documents folder and those who use another location.

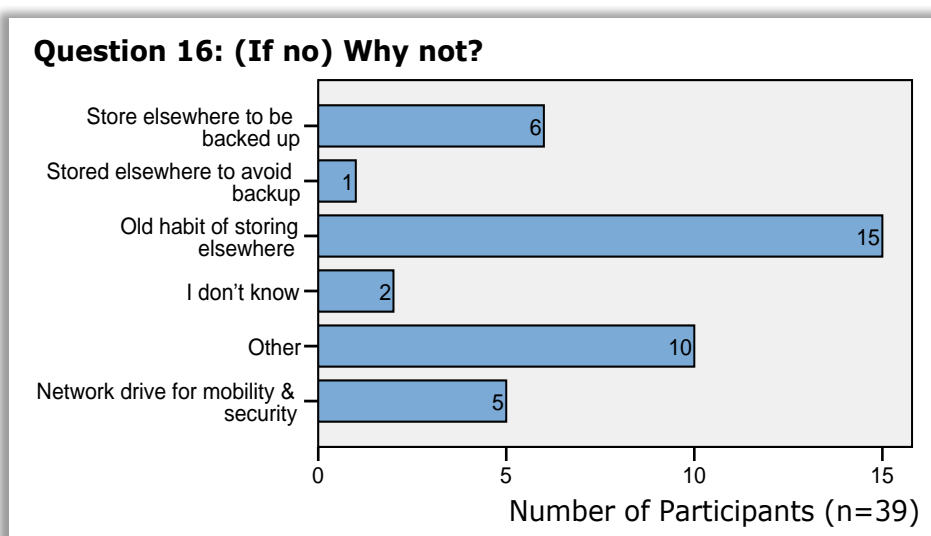


Figure 71: Question 16 - Bar chart showing why participants don't use the My Documents folder

15 respondents (38%) indicated that they stored their files elsewhere out of habit. A further 15 chose the 'Other' option for this question. Of those, 5 explained that they preferred using the network drives provided by the university, with most mentioning the wider availability of this drive at different locations, and several also mentioning that it is backed up. Another simply stated that they don't like to store anything in C drive.

One respondent who chose other didn't give a reason, and one said that it was a combination of the first three options.

Two responses related to a sense of ownership and control they felt was lacking with My Documents, with one saying *"I don't like to be told by MS where to put my documents"*, and another saying *"it's not my Documents unless I made that folder which I didn't"*. Another said it was because they *"need to organise them in customised fashion"*, perhaps referring to the fact that Windows XP and other applications create folders inside My Documents meaning it is not a fully customised location.

One person said they used Microsoft SharePoint in order to organise their files for sharing, and another person also mentioned using another location so that the files can be accessed by others. Another said *"I have set up my own structure next to it so it easy for me to back up and keep track of my stuff,"* and one simply (and ambiguously) said *"I have setup folders"*.

5.2.5 Survey Section 5: Creating and Naming Files

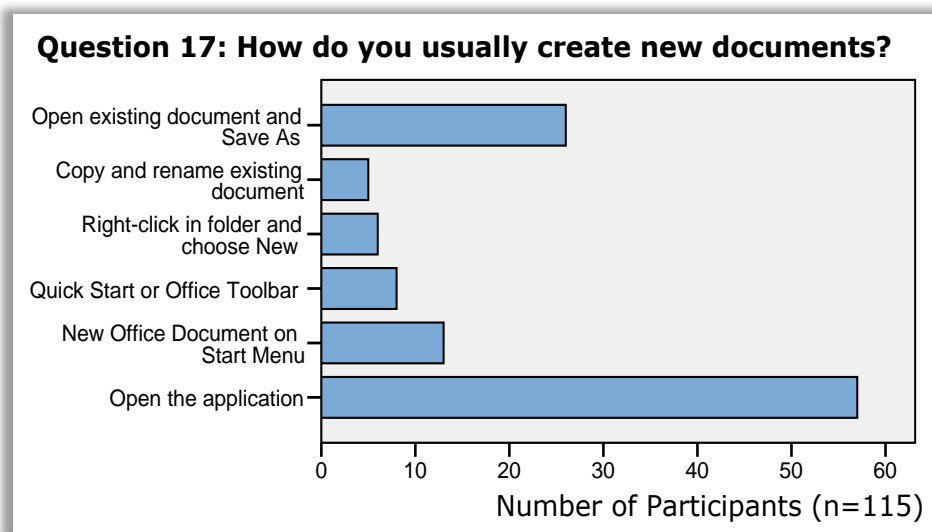


Figure 72: Question 17 - Bar chart showing how new documents are created

27% of respondents usually create new documents by making a copy of an existing one. The remaining participants use methods that create brand new blank documents, with the most common method (used by 50%) being to open the application they want to work with.

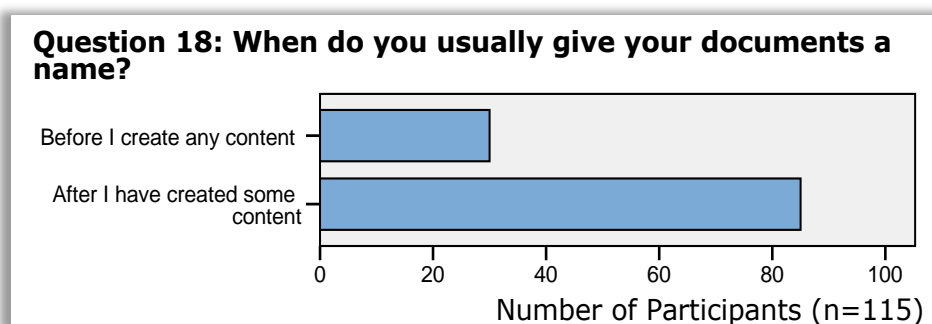


Figure 73: Question 18 - Bar chart showing when documents are usually named

26% of respondents usually name their documents before they start working on them, with the remainder working on the content first and only supplying a name afterwards.

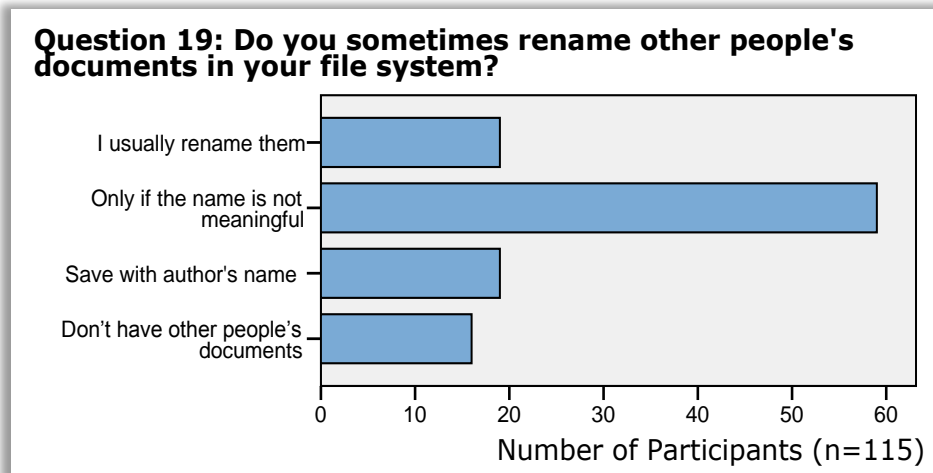


Figure 74: Question 19 - Bar chart showing whether other people's files get renamed

With regards to files received from other people, 17% of people will usually rename them, 17% of people usually won't. 14% say they don't normally have document created by other people, and the rest will rename sometimes if the original name doesn't make much sense.

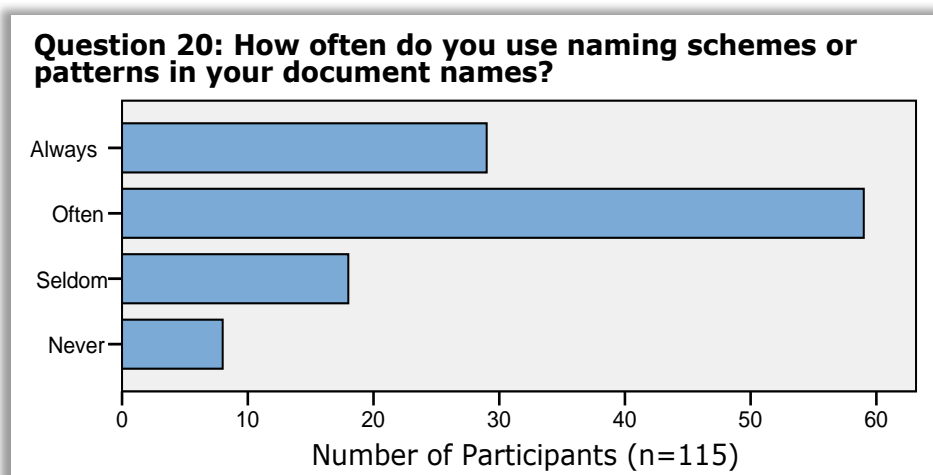


Figure 75: Question 20 - Bar chart showing how often document naming schemes are used

77% of respondents report always or often using naming schemes or patterns when naming their documents.

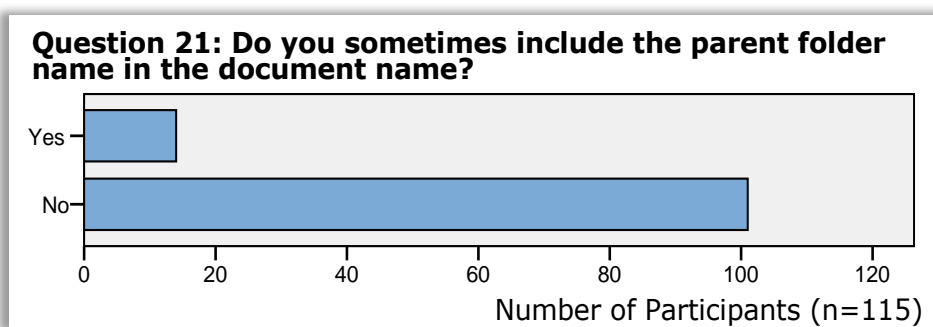


Figure 76: Question 21 - Bar chart showing whether the folder name is included in the file name

Only 12% of respondents say they sometimes include the name of the parent folders in the document name. This would indicate that the folder names provide additional metadata about the document that is not duplicated in the document name.

Question 22 asked the respondents who answered Yes to Question 21 to indicate why they include the parent folder name.

Six of the responses were related to it making it easier to find the file later, with comments such as: *"it's easier to find the document later (especially after a long time)"*, *"to identify the files easily"* and *"more information to search with"*. One person explained that it was *"to remind me of the document's source and when I have completed working on it I am reminded where it is to be filed."*

Two people noted that it was easier to track documents by course and year when you have multiple semesters of nearly identical information. One person mentioned it was because of the document's link to its parent folder. Another explained that it was due to the fact that when following shortcuts, *"it isn't always obvious where you are."* Finally, one noted it was simply habit.

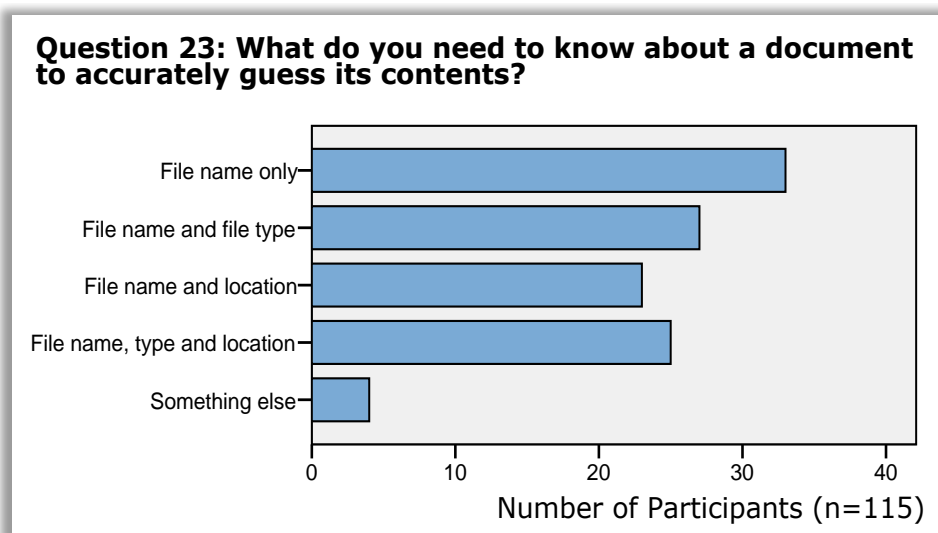


Figure 77: Question 23 - Bar chart showing minimum information needed to identify a document's contents

52% of respondents indicated that they could figure out what is in the file from just the file name and optionally type. 42% indicated that they would also need to know the location of the file in their folder structure, indicating that the folders are providing some key metadata they would need to be able to fully understand the file. 3 people didn't answer this question, and 4 indicated they need other information, with the following options being given:

- File name and file type and its location in my folders AND SIZE, DATE info
- File name and file type and date
- Document title, date last modified, person with last modification, version number
- File name, file type, location, and file size

So three people indicated that date would also be an important consideration for them, and two indicated that file size would be important.

5.2.6 Survey Section 6: Creating and Naming Folders

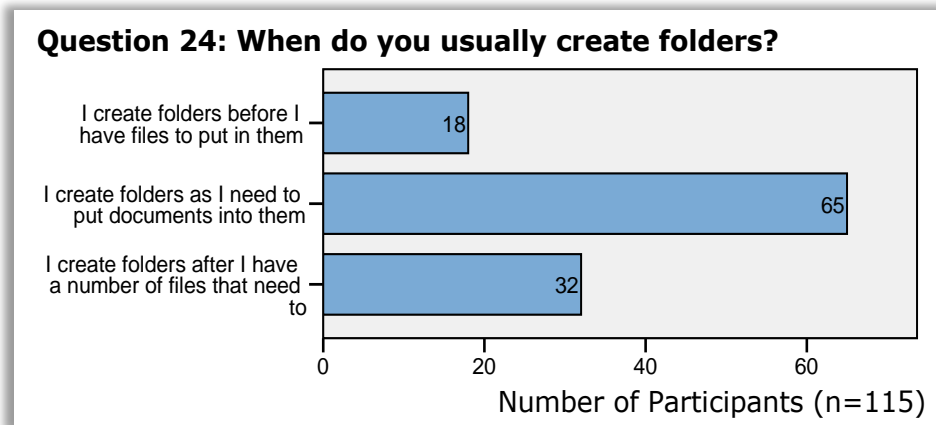


Figure 78: Question 24 - Bar graph showing when participants create folders

17% of respondents pre-arrange their folders in advance or use, 28% post-arrange after they have files that need to be organised, and the remainder create their folder on-demand. There is no significant correlation between when people create folders and when they name and save their documents.

Satisfaction differs depending on when respondents report creating folders ($f=2.99$, $\text{sig}=0.054$). Post hoc analysis shows that respondents who report creating folders only after they have files to put in them are less satisfied overall than respondents who create folder either in advance or as they need them.

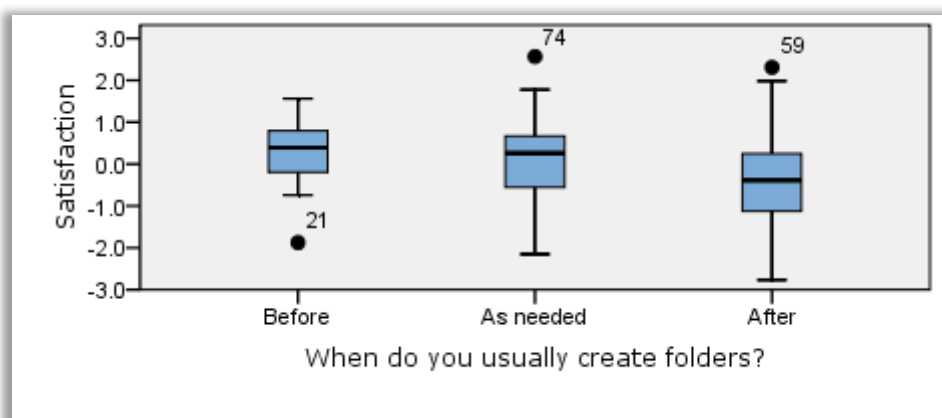


Figure 79: Box plot showing satisfaction is lower if folders are created after files

Question 25 asked the respondents to rate how important in their folder naming these items were: time periods, projects and courses, files types, subject and topic, purpose or use⁴.

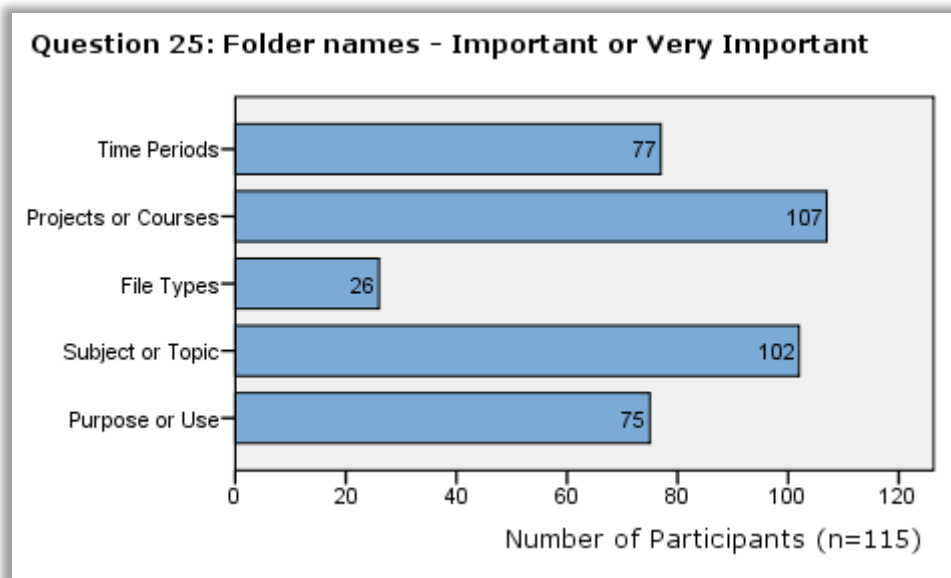


Figure 80: Question 25 - Bar graph showing the number of people who rated each folder naming item as either 'important or 'very important'

70% indicated that time periods were important or very important in their folder names and 90% indicated that projects or courses were important or very important in their folder names. Only 23% of respondents indicated that file types were important in their folder names. 89% of respondents felt that putting the subject or topic in their folder names was important or very important, and 65% of respondents felt that it was important to be able to organise their files by purpose or use.

5.2.7 Survey Section 7: Locating Documents

50% of respondents say they would know exactly where to find a recently used document, while 12% would have to browse for it. 33% of respondents indicate they would use one of the Recent Documents menus (either in Windows, or the application). Only 3% of respondents indicated they would use search in this situation. Of those that indicated 'Other', one explained "*I go to where I saved the file,*" which is essentially the same as the first response option. Another explained that "*the file will be on the desktop until filed. If it is filed I have an idea of is approximate location,*" which makes this a combination of the first two options. And the third said they'd use the Application's recent documents menu (option 3) but added "*why oh, why is it limited to 9*".

⁴ Although document genre was one of the dominant folder naming categories in the field study, it was not included as an option in the survey because few people understand what it means.

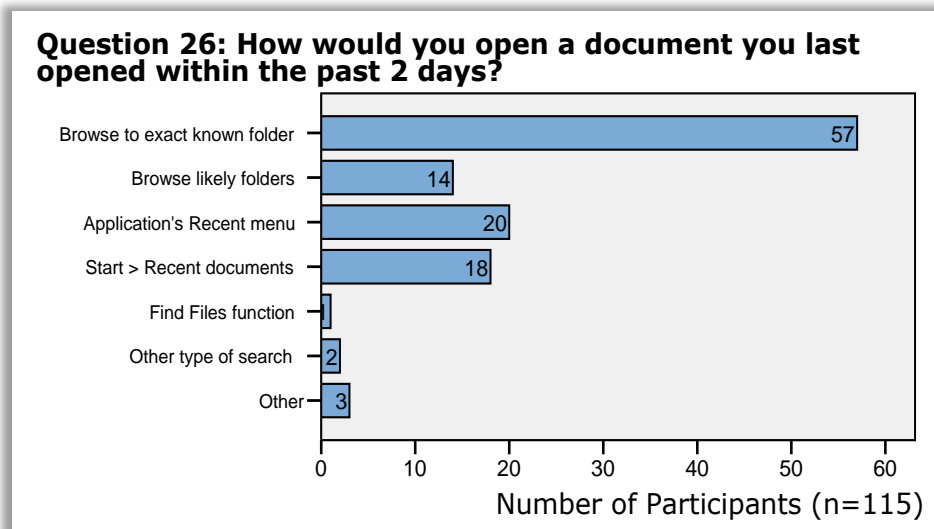


Figure 81: Question 26 - Bar graph showing find method for a recently-used document

There is a significant difference in satisfaction between each method ($f=4.27$, $\text{sig}=.001$). The participant who uses the Find Files function was extremely dissatisfied with their document management. This participant was removed from this test of differences between groups since only groups with two or more participants can be included in the analysis.

With this group removed, the difference between the groups was still significant ($f=3.42$, $\text{sig}=0.007$). Post hoc tests showed that people who report browsing to the exact known folder or using a Recent menu are significantly happier than people who browse likely folders or use some other method of locating the document.

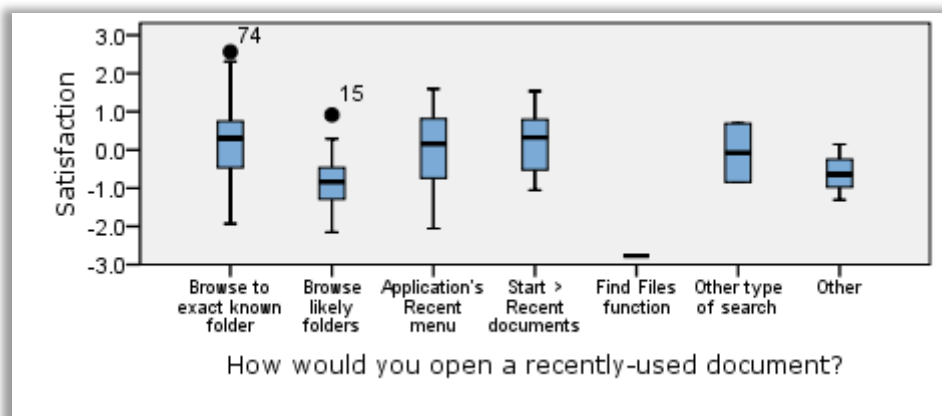


Figure 82: Box plot showing how satisfaction differs by find method for a recently-used document

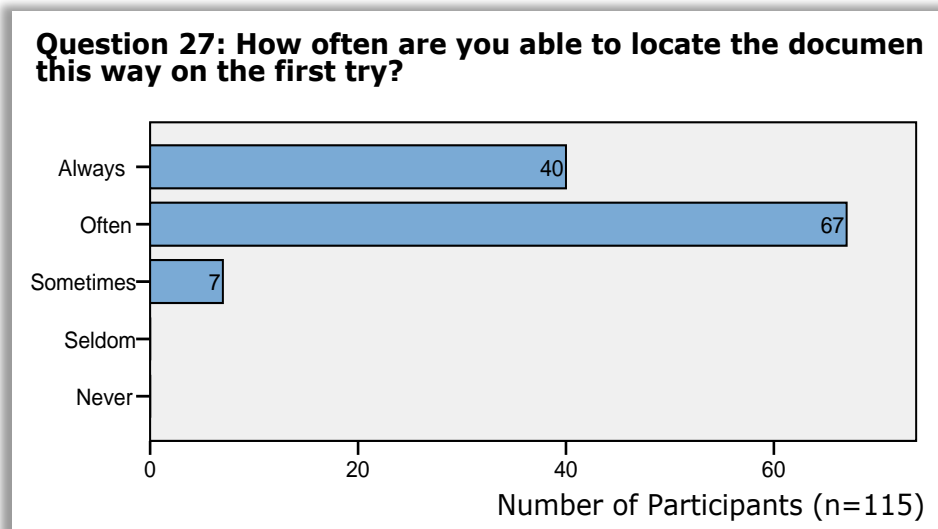


Figure 83: Question 27 - Bar graph showing find frequency for a recently-used document

Respondents report always or often finding their document on the first try 93% of the time. There is a significant difference in satisfaction between these three groups ($f=9.80$, $sig=.000$). People who can always locate the document are happier than those who often can, who in turn are happier than those who can locate it only sometimes.

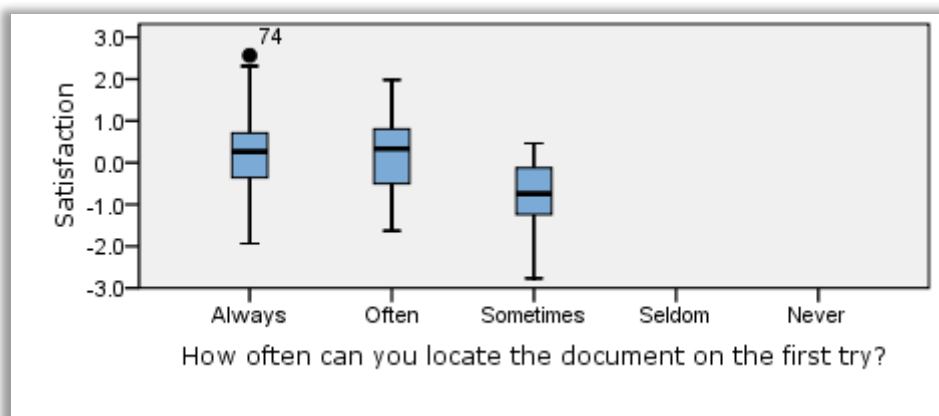


Figure 84: Box plot showing how satisfaction differs by search success frequency for a recently-used document

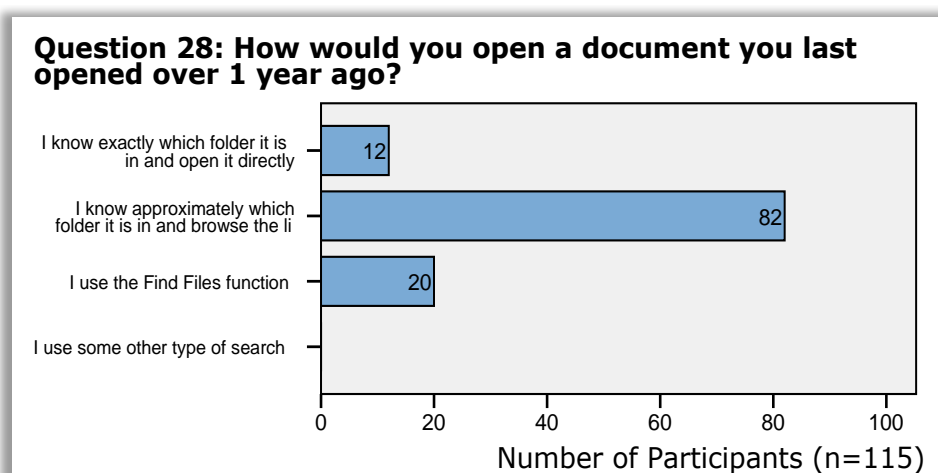


Figure 85: Question 28 - Bar graph showing find method for an old document

10% of respondents believe they would know exactly what folder a document was in after one year had elapsed. 17% say they would use a search function, and the remainder would browse until they recognised the file. There is a significant difference in satisfaction between each method ($f=4.57$, $sig=.012$). Those who know exactly what folder their document is in are significantly happier than those who need to browse or use a search facility.

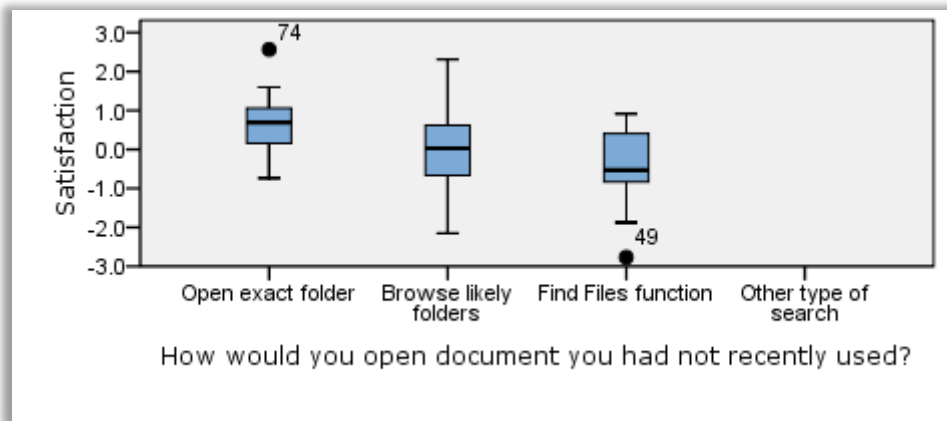


Figure 86: Box plot showing how satisfaction differs by find method for an old document

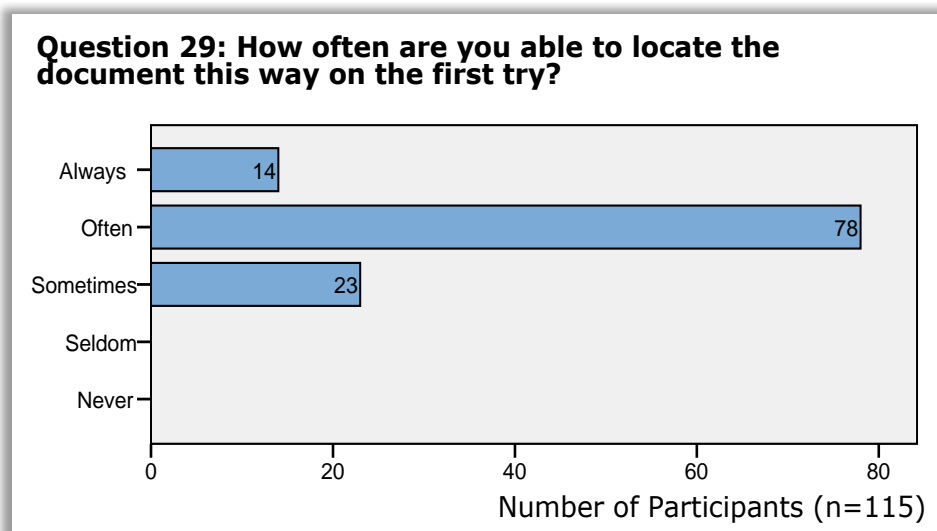


Figure 87: Question 29 - Bar graph showing find frequency for an old document

Respondents report always or often finding their document on the first try 80% of the time. As with the recent documents, participants are much happier if they can always or often locate the document on the first try and much less satisfied if they can only locate it this way sometimes.

5.2.8 Survey Section 8: Searching

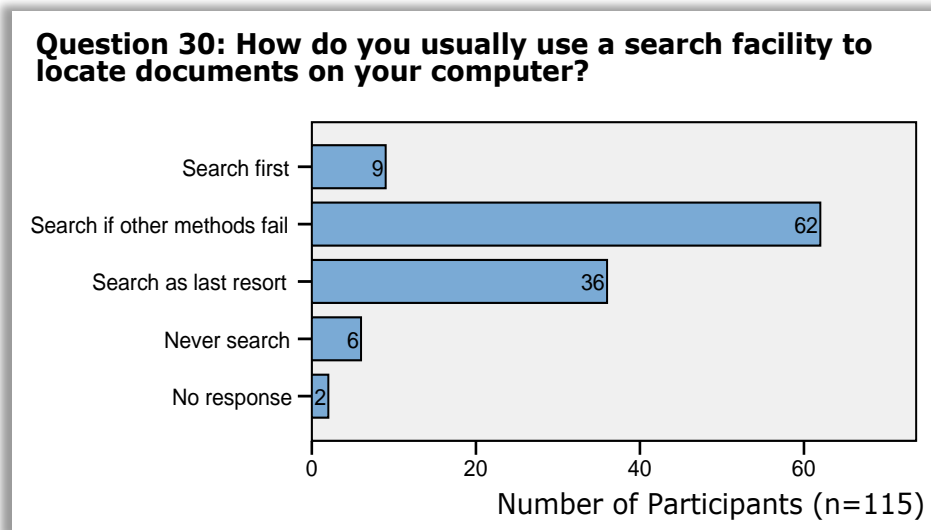


Figure 88: Question 30 - Bar graph showing how people use the search facility

8% of respondent report that search is their first strategy for finding documents. 54% will use search but only after trying other methods, and 31% will use it search only has a very last resort.

The following two questions were not answered by the 6 respondents who indicated they never search.

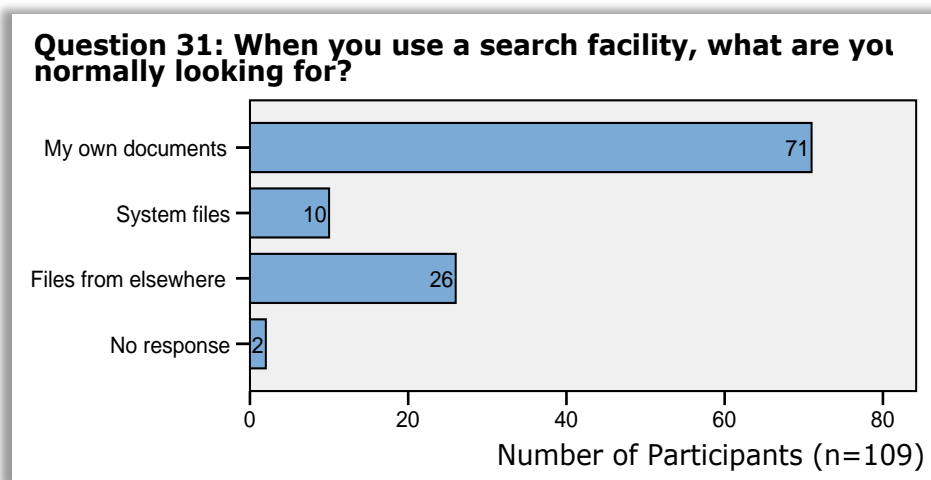


Figure 89: Question 31 - Bar graph showing most common search target

65% of respondent are looking for their own documents when they resort to using the search facility. 33% are looking for other people's files, either system files or files received from others or downloaded from the web. Those respondents who report they use a search facility to locate their own documents are less satisfied overall with their document management than participants that are looking for other people's documents ($f=2.75$, $sig=0.069$).

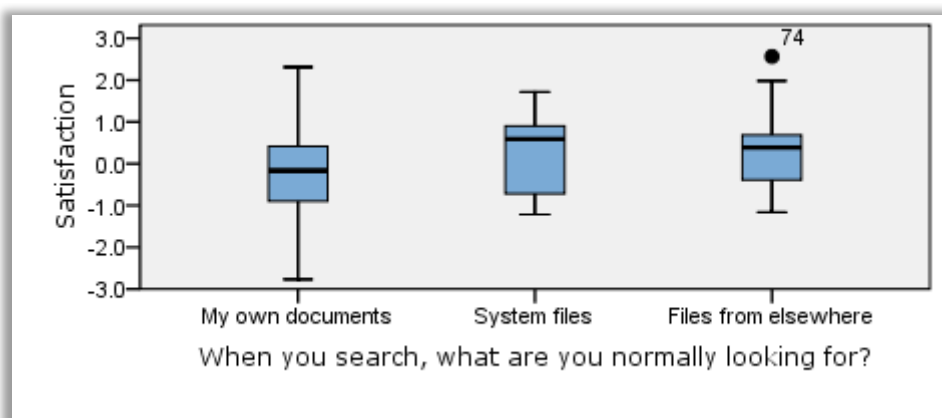


Figure 90: Box plot showing difference in satisfaction varying with target of search

Question 32 asks respondents how frequently they use these elements when searching: all or part of the file name, keywords or phrases from the file contents, file type or extension, file creation or modification date, and file size.

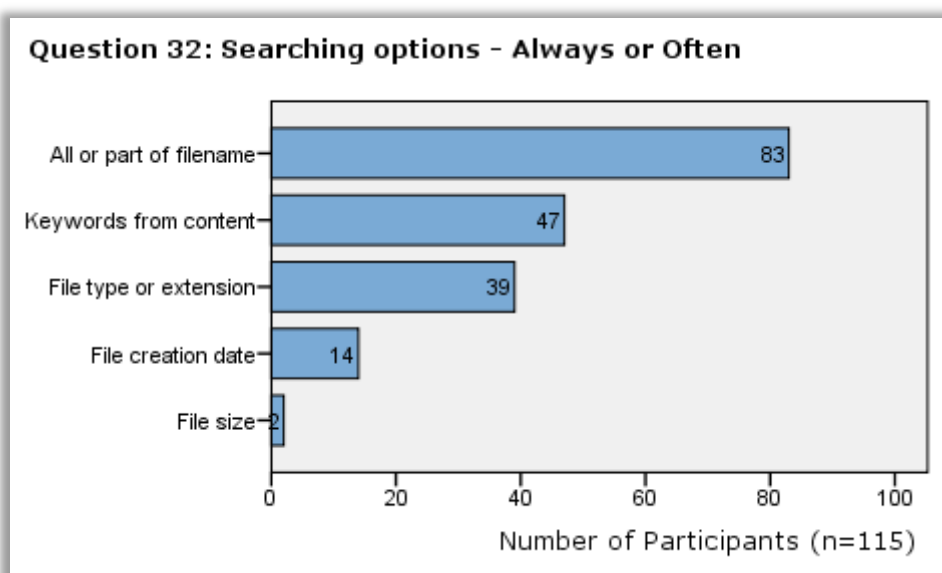


Figure 91: Bar graph showing the number of participants who indicate they always or often use each search criterion

76% of respondents use all or part of the file name always or often when they search, with 43% of respondents always or often search using keywords from the file, and 37% seldom or never using these. 36% of respondents reporting always or often using the file type or extension when they are searching, whereas 39% seldom or never use these options. 13% of respondents usually use a date in their searching, whereas 60% of people seldom or never do. Less than 2% of respondents often search using the file size, and 88% of people report they seldom or never use this option.

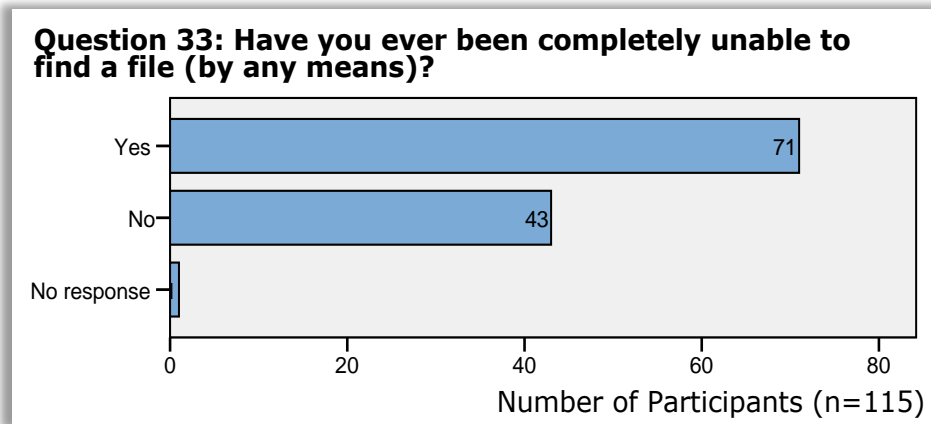


Figure 92: Question 33 - Bar graph showing how many people have failed to find a file

62% of respondents report that there has been at least one occasion where they have completely failed to find a file they were searching for. Those respondents are significantly less satisfied with their overall document management practices ($f=15.53$, $\text{sig}=.000$)

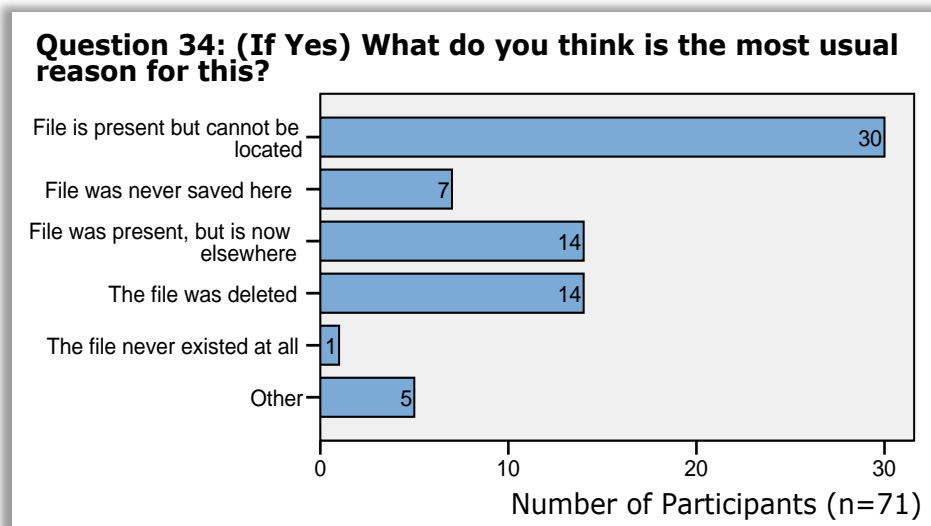


Figure 93: Question 34 – Bar graph showing common reasons for failing to find a file

Of the 71 respondents who had been unable to locate a file, 30 (42%) believed the file probably did exist but they just couldn't remember enough about it to locate it. Of those respondents who chose the 'Other' option, one said it is a combination of 2 and 4, that the file was either never saved to begin with or deleted. One said it was due to "hidden temp files", and another said it was due to someone else creating the doc and giving an uninformative file name. One singled out the "lousy" search function, and another said it could be any of the above: "if I knew the answer it wouldn't be lost!"

5.2.9 Survey Section 9: Viewing Documents

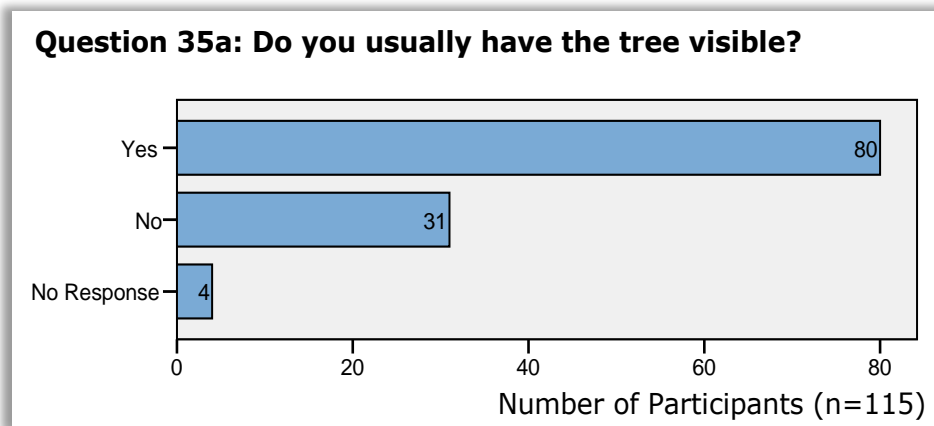


Figure 94: Question 35a - Bar graph showing how many participants view the tree

70% of respondents have the tree visible when viewing their files and folders.

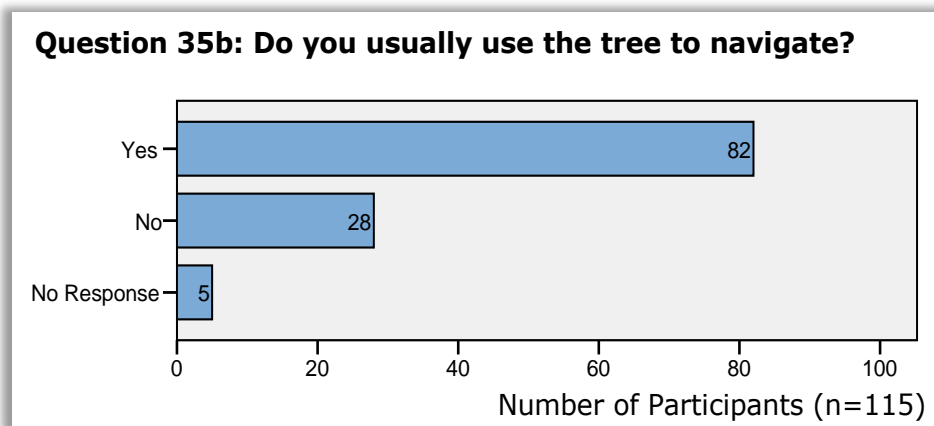


Figure 95: Question 35b - Bar graph showing how many participants use the tree to navigate

71% of people indicate that they use the tree to navigate. Interestingly, there are two respondents who say they usually use the tree to navigate, but say they do not usually have the tree visible.

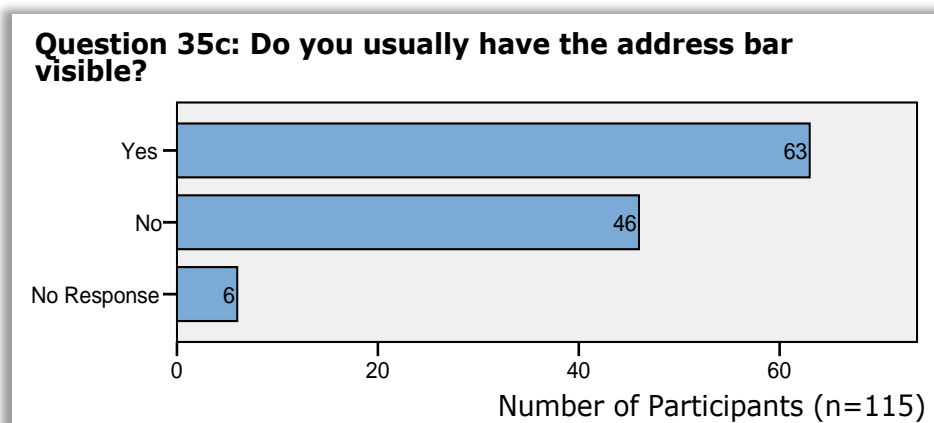


Figure 96: Question 35c - Bar graph showing how many participants have the address bar visible

Respondents are split in whether they usually have the address bar visible, with 55% having it visible and 40% having it hidden.

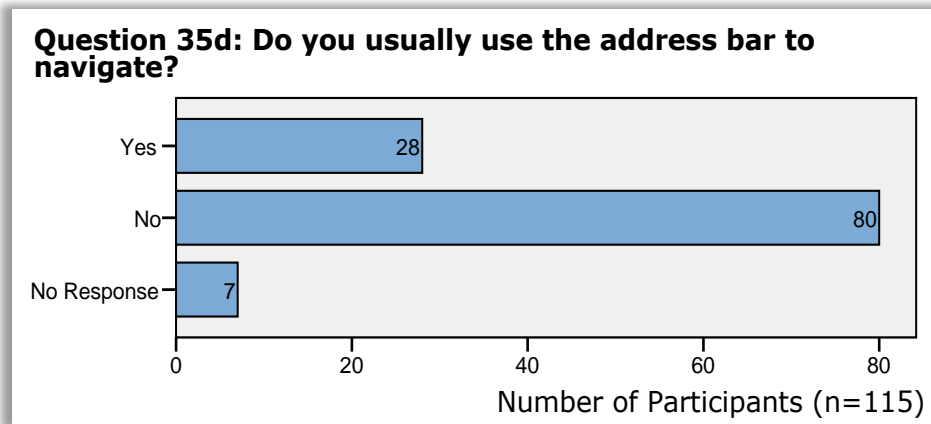


Figure 97: Question 35d - Bar graph showing how many participants use the address bar to navigate

70% of respondents do not use the address bar to navigate.

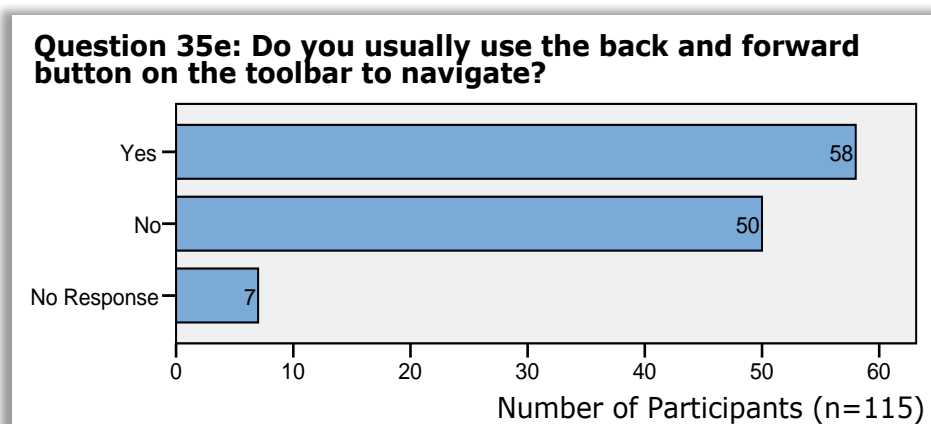


Figure 98: Question 35e - Bar graph showing how many participants use the back and forward buttons to navigate

50% of people use the back and forward buttons when navigating, whereas 44% do not.

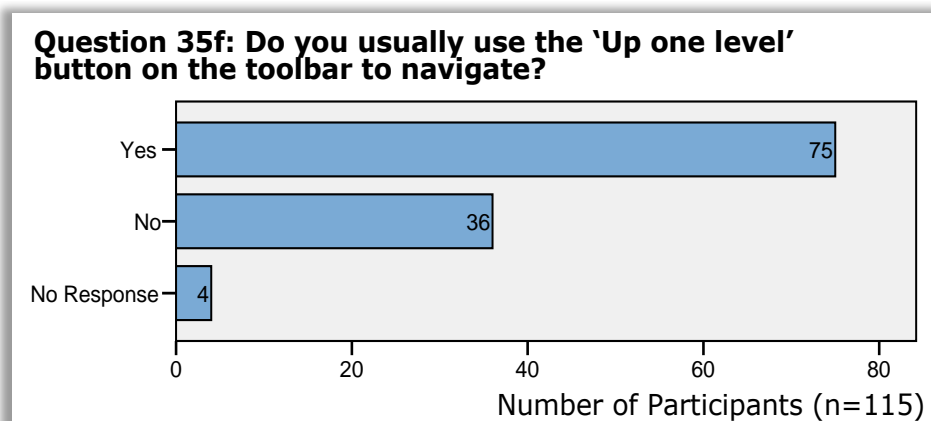


Figure 99: Question 35f - Bar graph showing how many participants use the 'Up One Level' button on the toolbar to navigate

65% of respondents use the 'Up one level' button to navigate up the tree.

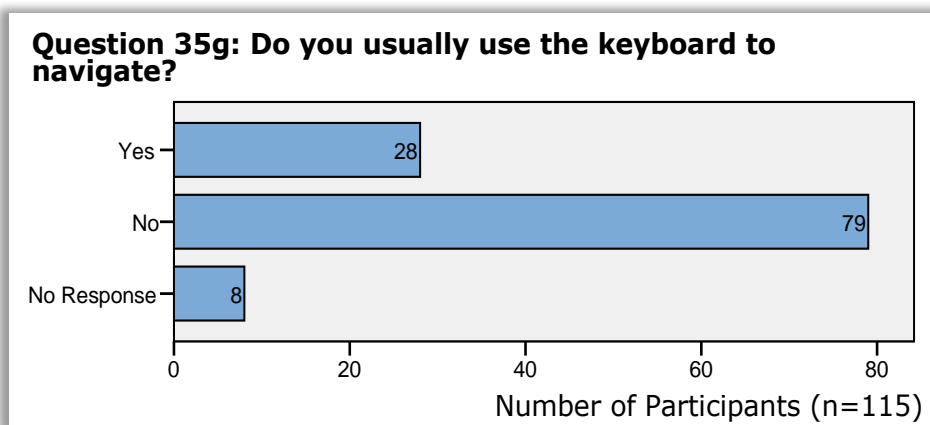


Figure 100: Question 35g - Bar graph showing how many participants use the keyboard to navigate

Most people (69%) report navigating with the mouse rather than the keyboard.

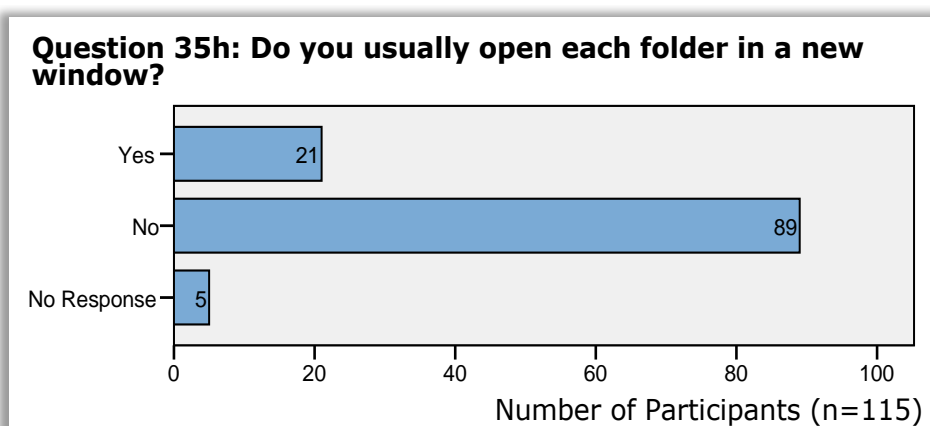


Figure 101: Question 35h - Bar graph showing how many participants open each folder in a new window

77% of respondents open folders in the same window, as opposed to opening a new window for each folder. Those respondents who use the same window are significantly more satisfied on average than the respondents who open new windows ($f=4.24$, $\text{sig}=0.017$). A Chi-square test indicates that people who open folders in a new window are less likely to use the tree to navigate (Chi-square value=78.942, $\text{sig}=0.000$).

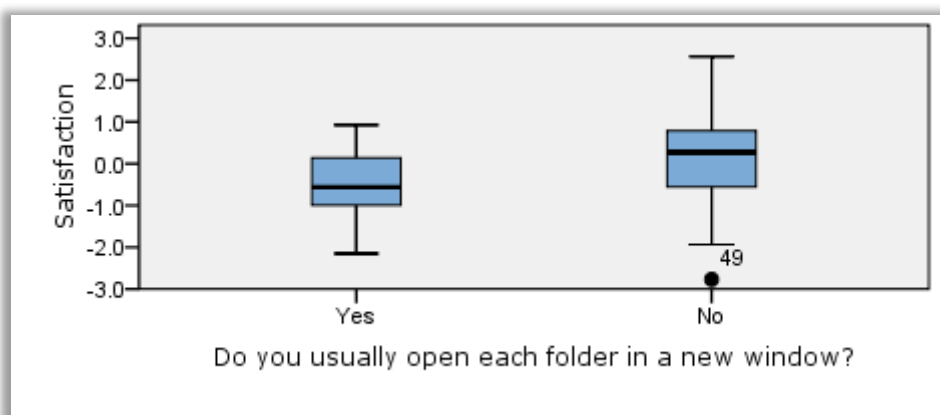


Figure 102: Box plot showing respondents are less satisfied if they open each folder in a new window

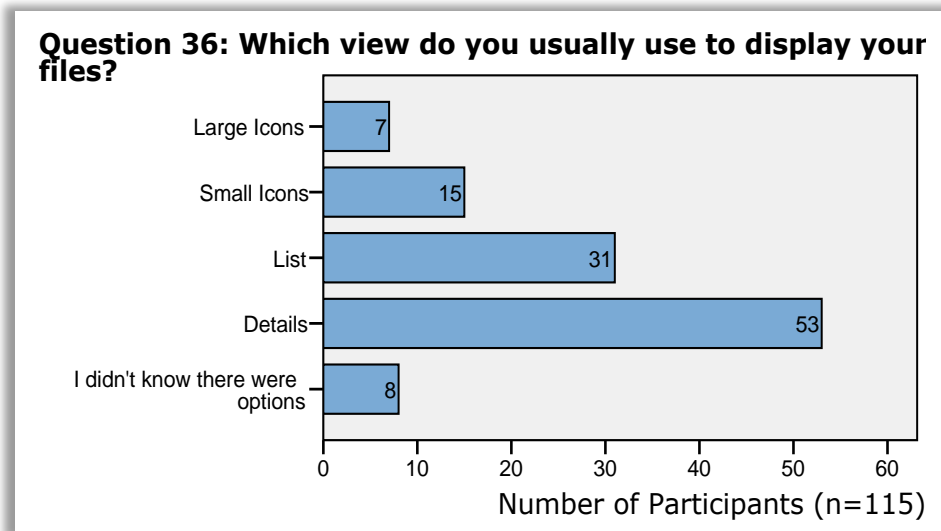


Figure 103: Question 36 - Bar graph showing most common folder view

The majority of people (53%) use one of the icons views (Large Icons, Small Icons, List), while 49% of people use the details view which offers more information. 8 respondents use the default List view because they didn't know there were options to change the view.

The 53 people who use the details view were asked what sorting options they use most frequently.

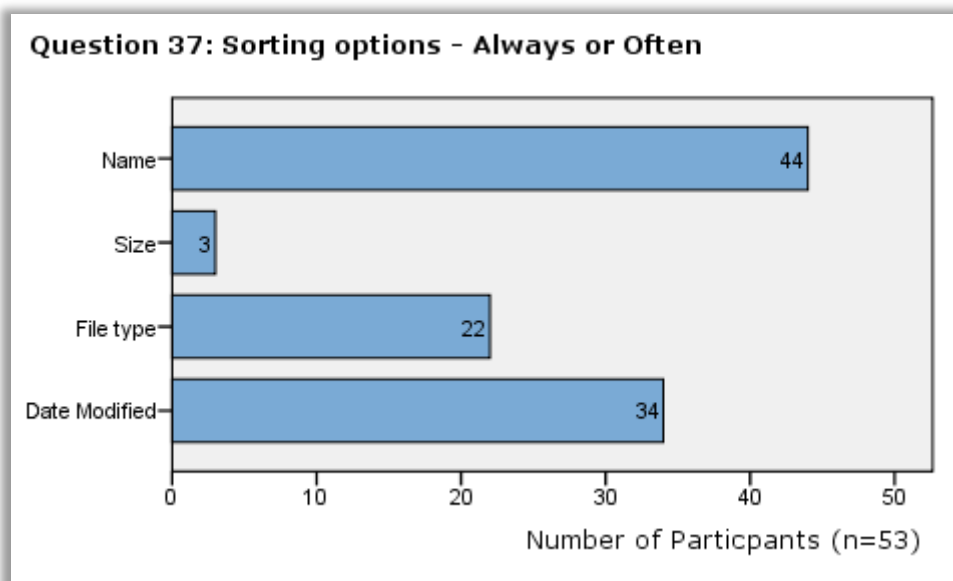


Figure 104: Question 37 - Bar graph showing how many participants always or often participants sort by each criterion

Name was the most commonly used sort option, with 83% of respondents saying their documents were sorted by name always or often. File size was the least frequently used option with only 6% of respondents reported usually sorting by size. File type was frequently used by 41% of respondents. Date modified was used frequently by 64% of respondents.

5.2.10 Survey Section 10: Versions

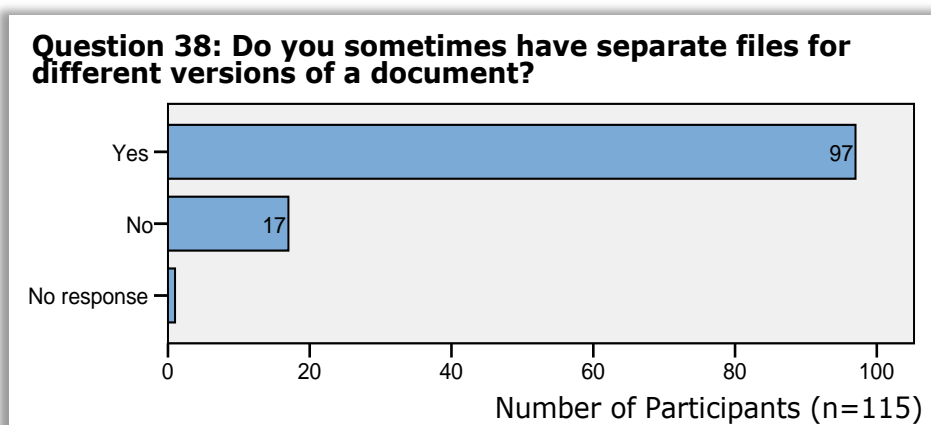


Figure 105: Question 38 - Bar graph showing how many participants have separate files for document versions

83% of respondents have multiple versions of a document stored as separate files. The following questions apply only to the 97 respondents who do have document versions in separate files.

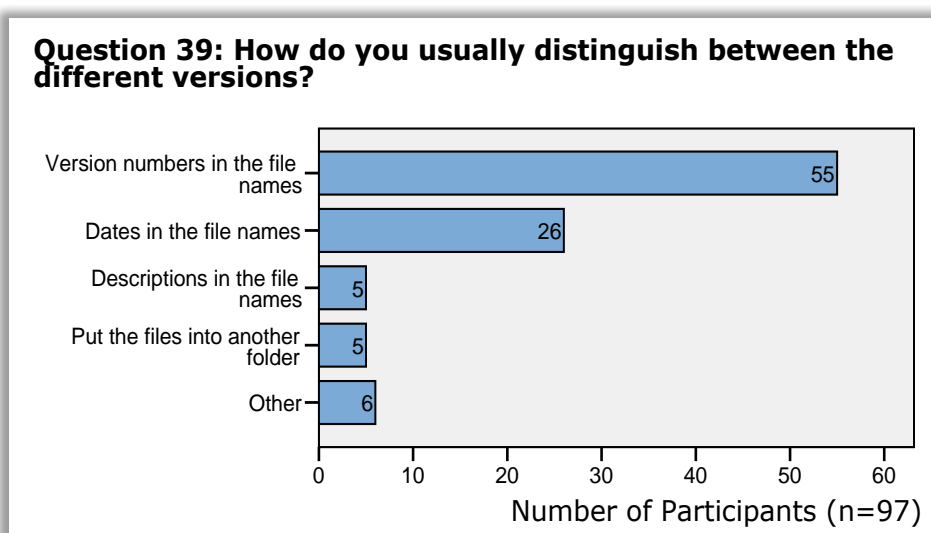


Figure 106: Question 39 - Bar graph showing how participants distinguish between file versions

Of those 97 participants who have multiple files for version of documents, the majority (57%) distinguish between them using version numbers in the file name. The next most common option is to use dates to differentiate the version, with descriptions or folders being used by only 5 respondents each.

The 6 respondents who chose 'Other' all indicated that they would use a combination of these methods. 2 people said they use all of these methods, a further 2 says they use all of the first three, 1 person uses the first two, and 1 uses the first two plus also the name of the person who changed the file.

In addition, one respondent who chose dates added the extra information that they used either dates or semesters. And one respondent who said they put the files in another folder added "I actually

combine the above and put the version name in the file name AND put them in a separate folder (usually labelled 'Old Whatever')"

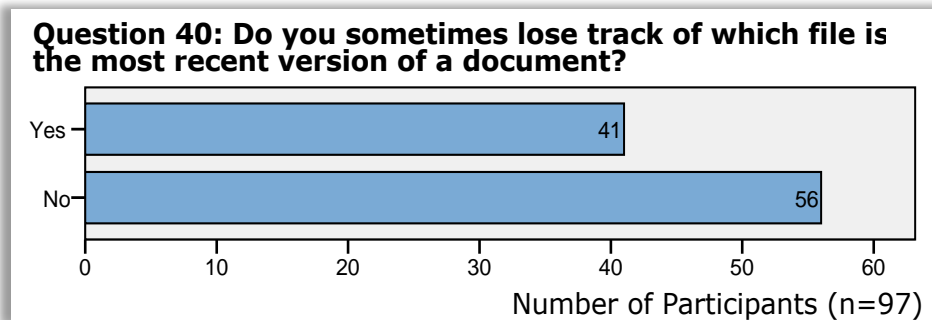


Figure 107: Question 40 - Bar graph showing how many participants lose track of which file is the most recent version of a document

The 42% of respondents who report they sometimes lose track of which document is the most recent version were significantly less satisfied with their document management overall ($f=22.11$, $sig=.000$).

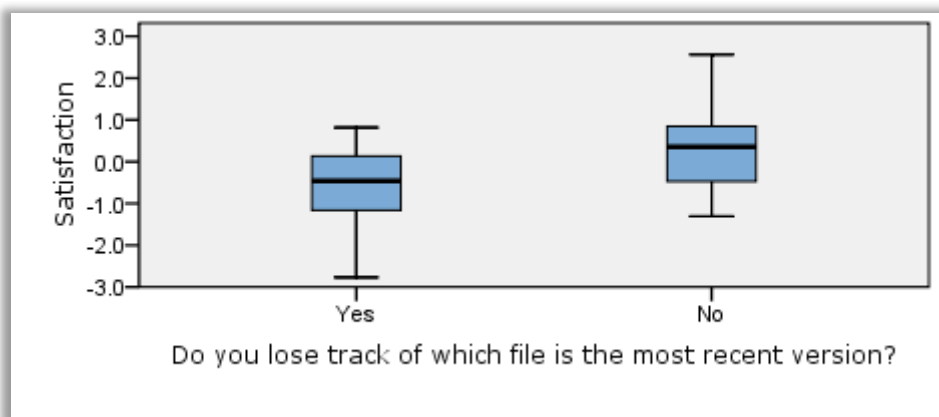


Figure 108: Box plot showing people who sometimes lose track of the most recent version of a document are less satisfied overall

Question 41 asked the people who sometimes lose track of which file is the current version why they think this happens. 37 of the 41 people responded.

The most common reason, reported by nine people was that they have no systematic way of identifying versions, or at least, no consistent way of doing it. One mentioned not being able to store dates in the name of the file, which is due to the slashes in the most common short date format being prohibited in file names. Five people didn't specifically mention that it was a systematic problem, but just said that their files weren't named well enough for them to be able to tell the version from the file name.

The next most common reason was that they forget to add the version identifiers, with a couple noting that it is particularly a problem when you come back to a document after a few days. On the same theme, some mentioned that they were too busy or too lazy or just didn't take the time to assign proper identifiers, and two simply said it was carelessness

Four people identified collaboration as a problem, saying that other people gave documents different names, used different version identifiers or forgot to change version identifiers.

Three people fingered multiple locations as the problem, saying for instance that they “*transfer documents between computers and forget where I last worked on the file.*” Three report running into problems because different versions are stored in different folders on the same computer, which can cause them to overlook the most recent one.

One person didn’t know, one said they mistakenly put the wrong version numbers on a file, one person mentioned problems with email attachments being saved in temporary locations, and another said they had accidentally overwritten a version of a file.

5.2.11 Survey Section 11: Copies

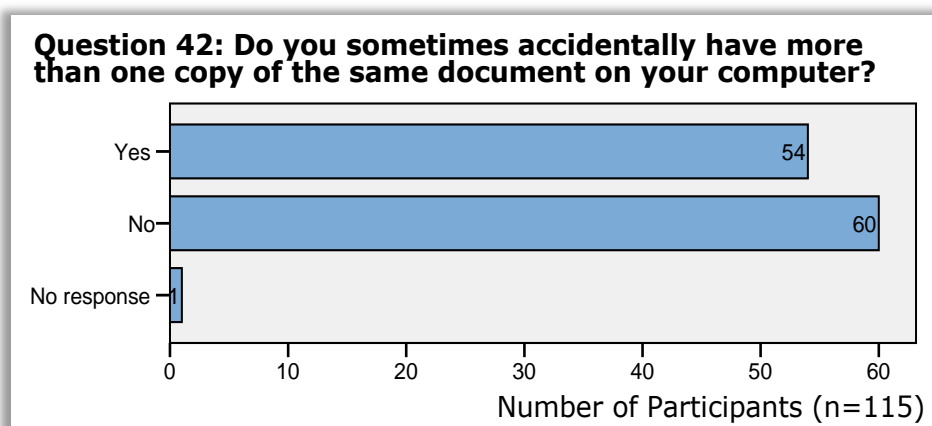


Figure 109: Question 42 - Bar graph showing how many people accidentally have multiple copies of the same document

47% of respondents report sometimes accidentally having more than one copy of the same document. These people who duplicate documents are overall significantly less satisfied with their document management practices overall ($f=16.66$, $\text{sig}=.000$).

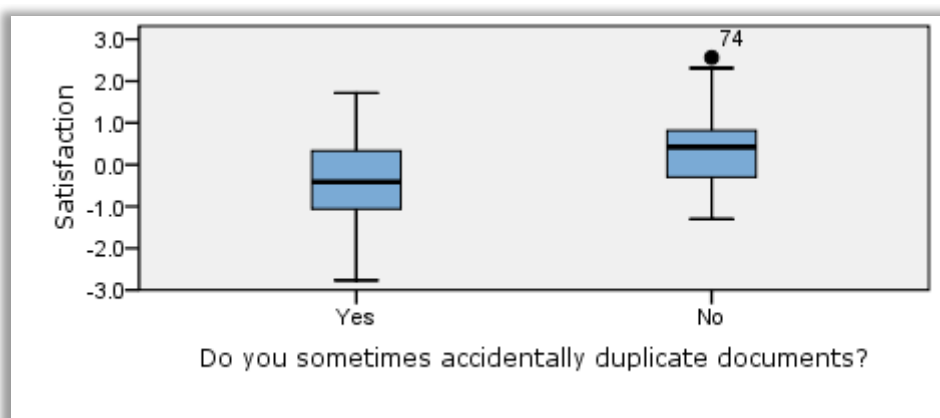


Figure 110: Box plot showing people who report accidentally duplicating documents are less satisfied overall

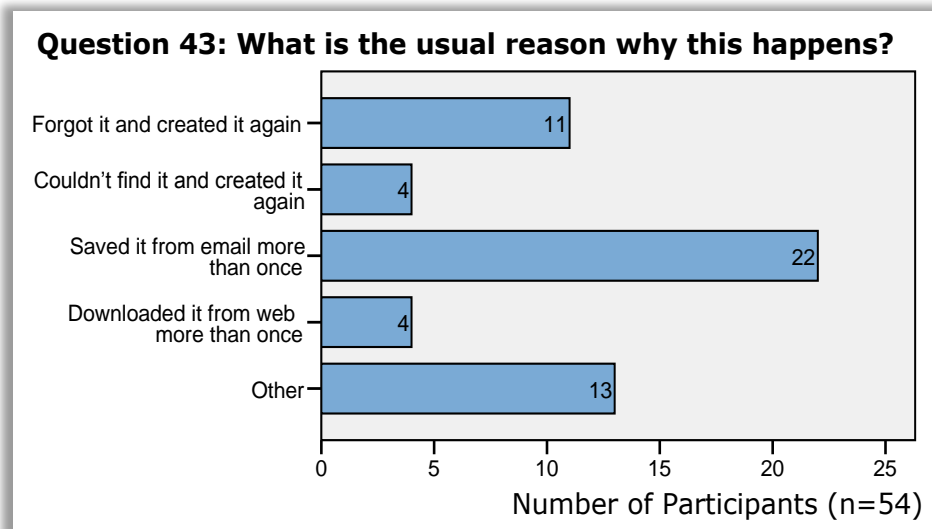


Figure 111: Question 43 - Bar graph showing how accidental duplication happens

The most common reason people think they have multiple documents is due to saving email attachments multiple times. However, forgetting a document existed and recreating it was considered the most common reason by 20% of respondents.

Of the 13 people who answered 'Other', four say they have duplicated files deliberately. Two say they have this situation because of having backup copies of a file. Two mentioned copying files to a different location to make them easier to upload into the university's learning management system Cecil. One person said that all the options apply, three said it arose because of transferring files between computers, and 1 said they thought they accidentally save the same file with a different name.

5.2.12 Survey Section 12: Deleting and Backup

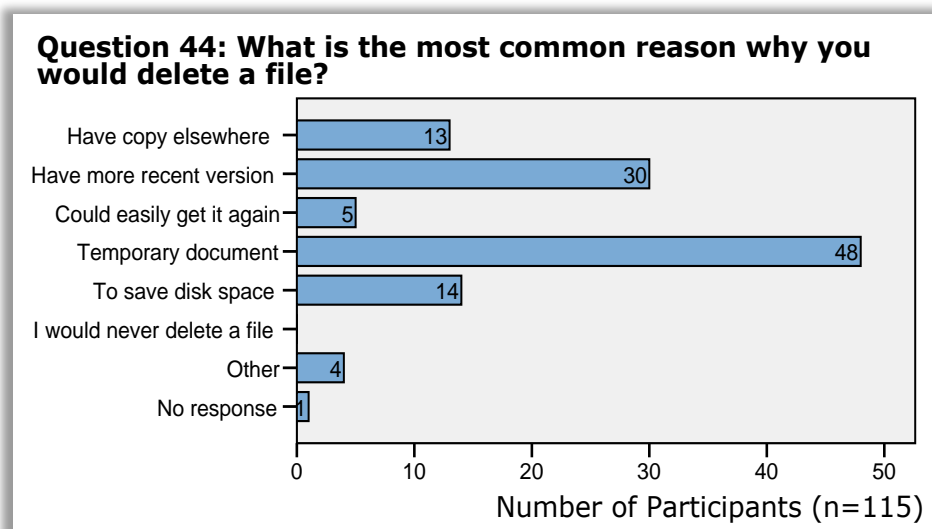


Figure 112: Question 44 - Bar graph showing reasons for deleting a file

48 people (42%) say the most common reason for deleting a file was that the file was temporary in the first place. This creation of temporary files is something that needs to be investigated in more

detail. The second most common reason (26% of respondents) is deleting files that embody older versions of another file. One respondent who answered other indicated that it is a mixture of both of those two reasons. Another respondent said that it's a mixture of all of the above. The other two indicated that they would delete the file if they couldn't foresee any future use for it.

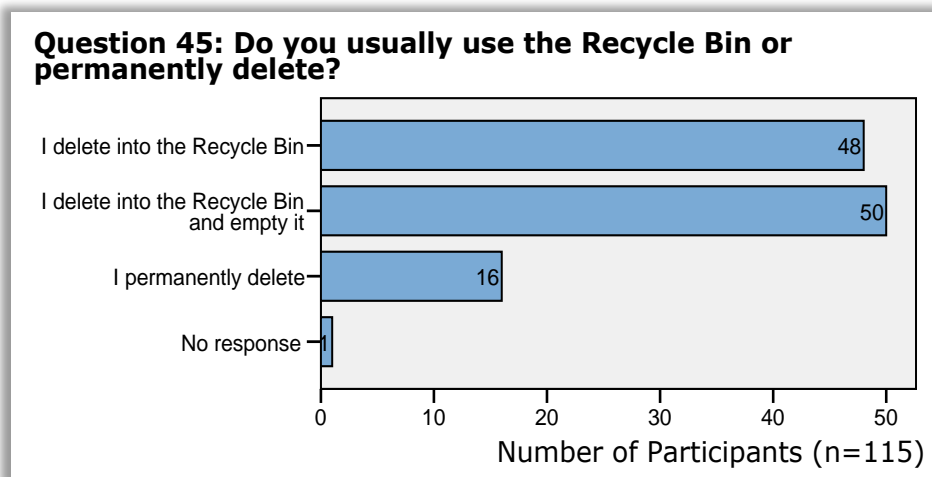


Figure 113: Question 45 - Bar graph showing use of Recycle Bin

Only 42% of respondents actually use the recovery function of the Recycle Bin. 14% permanently delete, and 44% use the Recycle Bin but frequently empty it, making the documents in it unavailable for recovery.

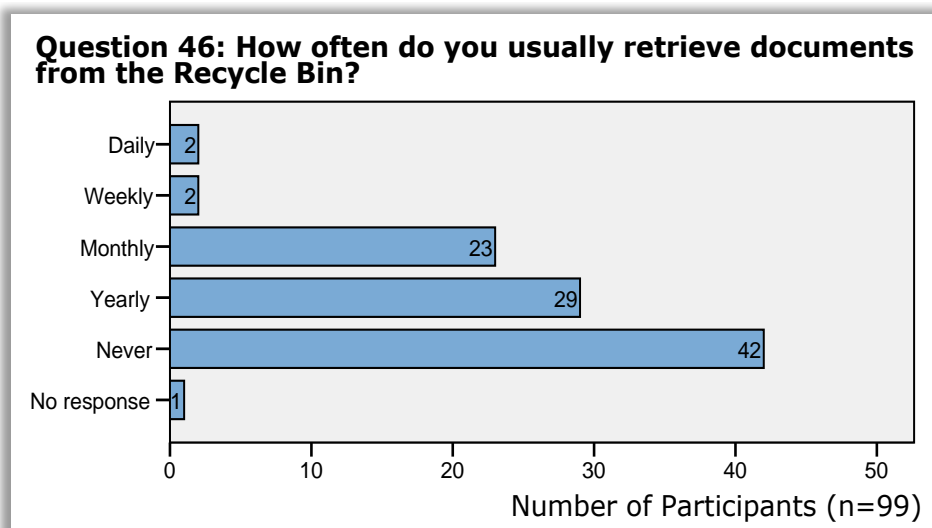


Figure 114: Question 46 - Bar graph showing how often documents are received from the Recycle Bin

The majority (42%) of respondents who use the Recycle Bin report that they never retrieve deleted items from it. Most users report very infrequently retrieving items, on a monthly or yearly timeframe. A small minority of users report retrieving items from the Recycle Bin daily or weekly.

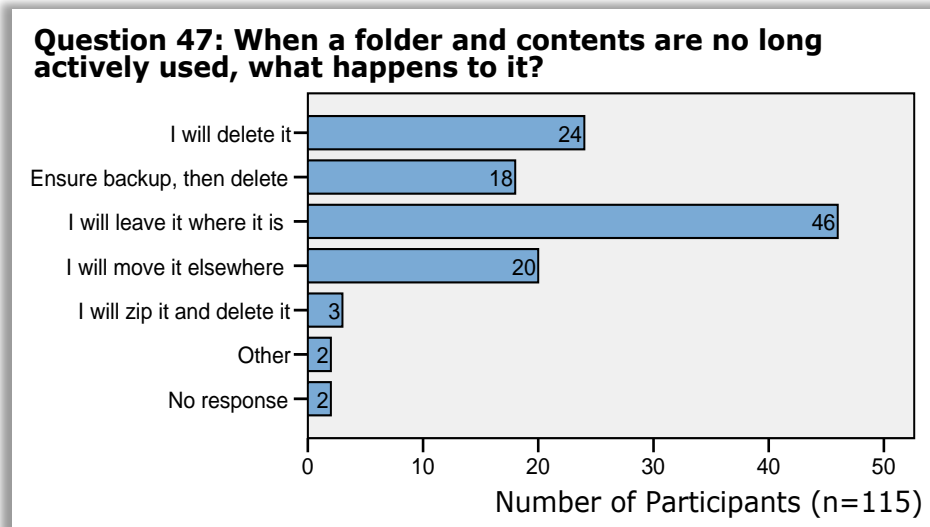


Figure 115: Question 47 - Bar graph showing the final destination for inactive documents

Most people (40%) don't take any special action on folders that are no longer being actively used. 39% will delete folders, with most indicating they will make sure a backup is available first. 17% will move the folder to another location.

There is an overall difference in satisfaction between these groups ($f=2.379$, $\text{sig}=0.042$). Post hoc tests (LSD) indicate that respondents who report ensuring they have a backup then deleting are happier than those who leave it where it is ($\text{sig}=0.002$) and happier than those who move it to another location ($\text{sig}=0.005$).

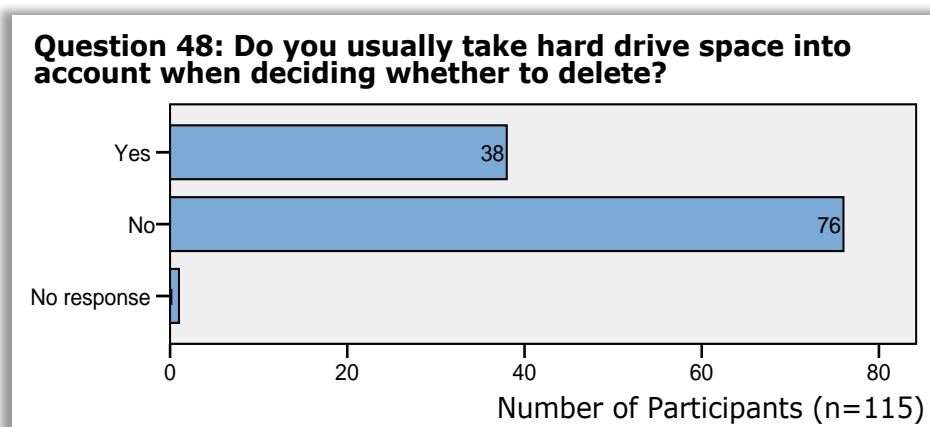


Figure 116: Question 48 - Bar graph showing how often hard drive space is considered in delete decisions

The majority of respondents (66%) report they do not consider hard drive space when they decide whether to delete files or not.

5.2.13 Survey Section 13: Demographics

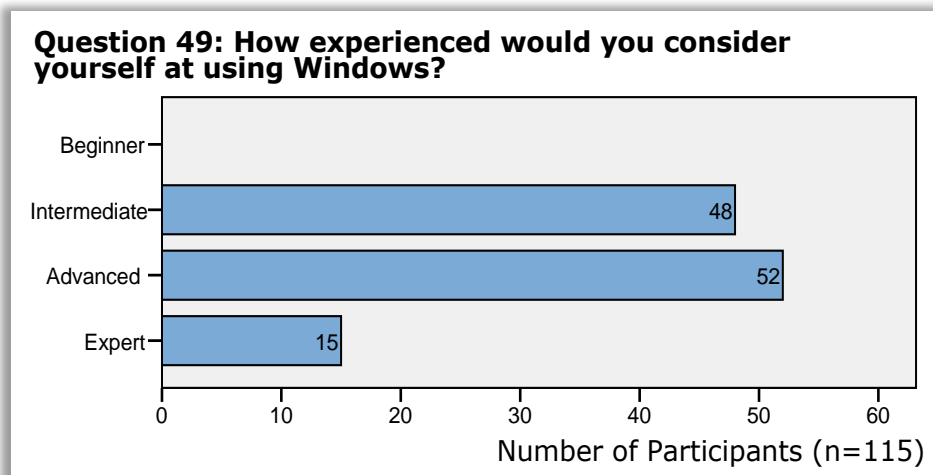


Figure 117: Question 49 - Bar graph showing self-reported Windows experience

None of the survey participants consider themselves to be beginners at using Windows. 13% report themselves to be experts, with the remainder fairly evenly split between Intermediate and Advanced.

There is a difference in how satisfied each of these groups is with their document management practices ($f=2.82$, $\text{sig}=0.064$). People who consider themselves Intermediate are less satisfied overall than those who consider themselves Advanced ($\text{sig}=0.026$). Those who considered themselves Experts had a similar average satisfaction to the Advanced users, but due to the lower sample size, this difference was not statistically significant.

Experts and Advanced users may differ from Intermediates in the way they use Windows, and this may have a bearing on how successful their document management practices are. A Chi-Square test was performed to test whether there was a relationship between the self-reported level of experience and each of the windows use items in Question 35.

There was a significant relationship between experience and reported use of the tree for navigation (Chi-Square value=9.557, $\text{sig}=0.008$). Advanced and Expert users make more use of the tree for navigation than do Intermediate users.

The other significant relationship was between experience and whether or not the respondent reports keeping the address bar visible in Windows Explorer (Chi-Square value=11.788, $\text{sig}=0.003$). Again, Advanced and Expert users are disproportionately more likely to have the address bar visible than Intermediate users.

No other questions about document management or Windows features showed any significant relationship with self-reported Windows proficiency.

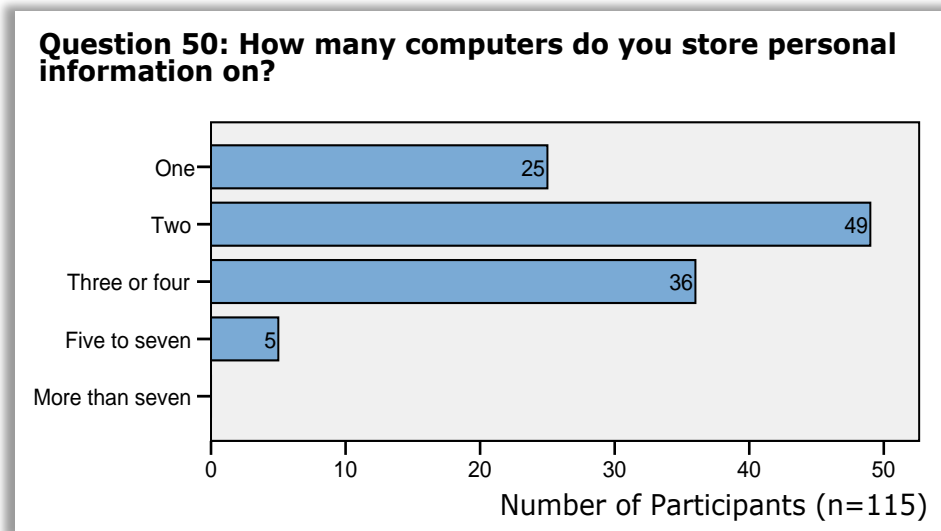


Figure 118: Question 50 - Bar graph showing number of computers used

Most people report storing personal information on more than one computer, with only 22% using only one computer.

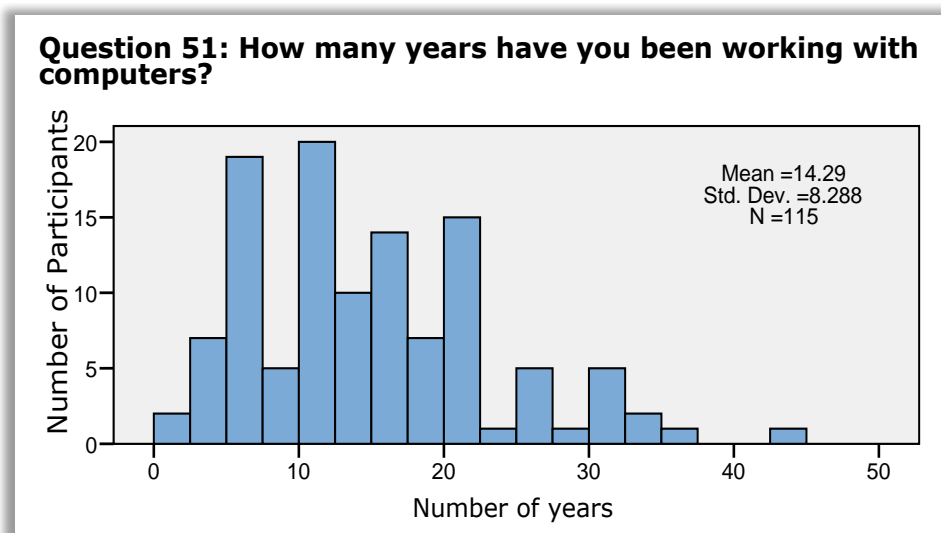


Figure 119: Question 51 - Histogram showing years of computer use

The mean number of years people have been working with computers is 14, and a standard deviation of 8.

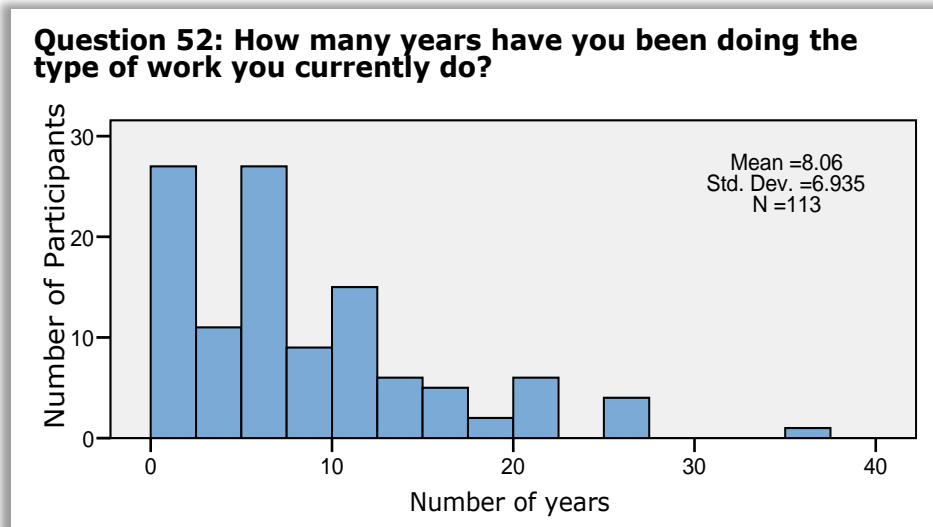


Figure 120: Question 52 - Histogram showing years in current line of work

A high number of participants have been in their current job only a couple of years. The mean number of years in the same type of work is 8 years.

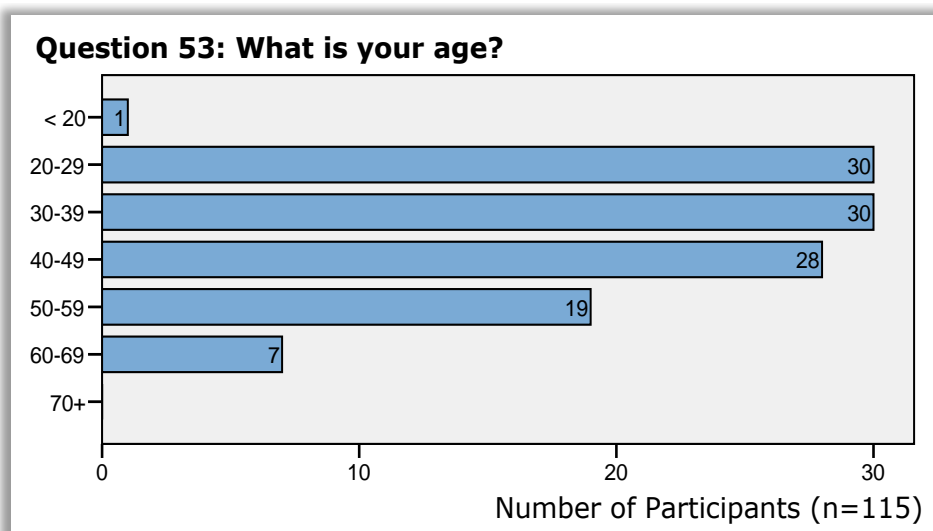


Figure 121: Question 53 - Bar graph showing age distribution

There was a wide range of ages represented in the survey respondents, with one participant under 20, seven in their sixties and the rest fairly evenly spread.

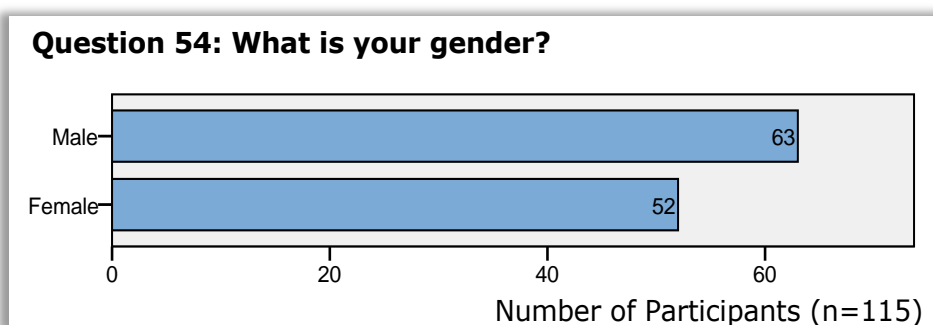


Figure 122: Question 54 - Bar graph showing gender

The respondents were 55% male and 45% female.

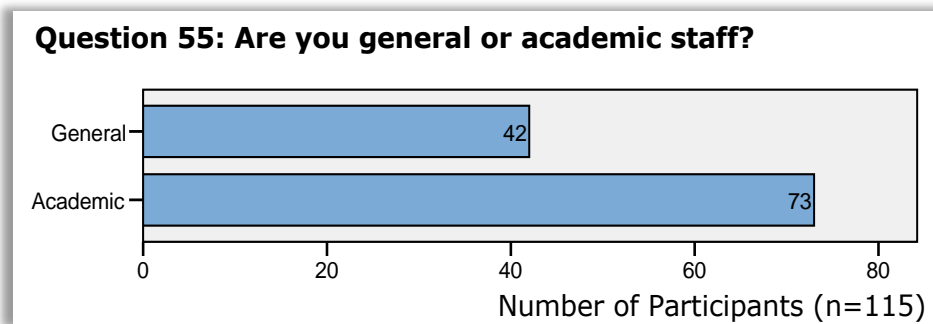


Figure 123: Question 55 - Bar graph showing proportion of Academic and General staff

63% of respondents were academic staff, with the remainder being general staff.

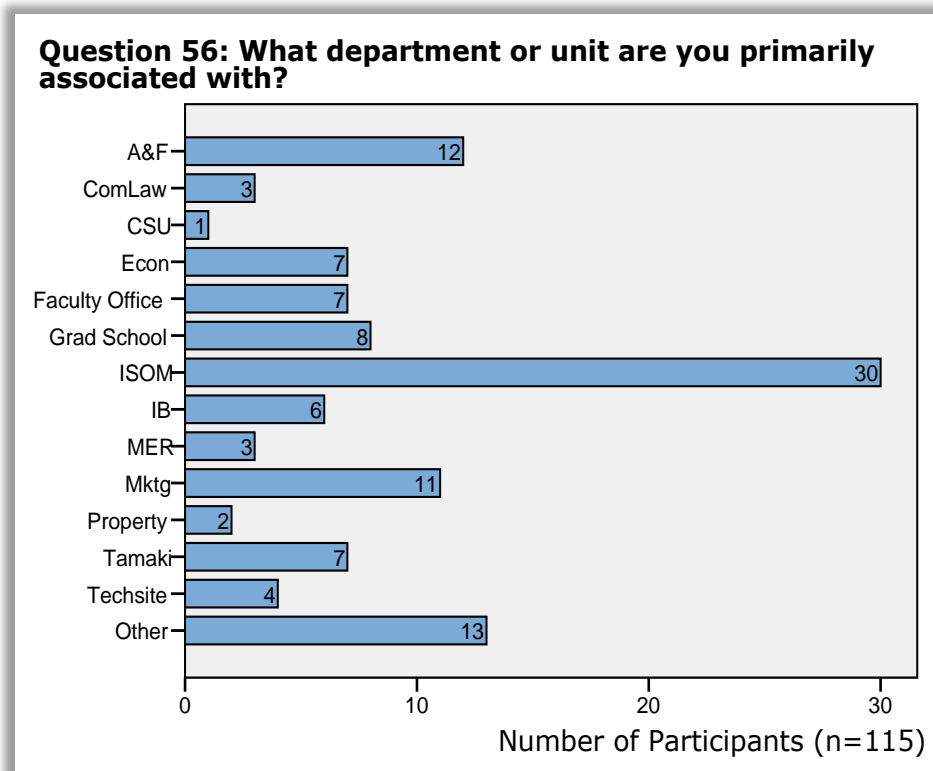


Figure 124: Question 56 - Bar graph showing proportion of participants in each department

The majority of respondents (26%) are in the ISOM department. This is likely to be due to a higher perceived obligation to respond because these people know the researcher personally. All other departments and units within the Business School are represented in the sample.

Question 57 asked for people's job titles. A very wide variety of responses was received, including all the University of Auckland's standard academic roles. The most common responses were Lecturer (16%) and Senior Lecturer (7%).

5.2.14 Survey Section 14: Comments

The final section of the survey was simply a single free text box asking the participants if they had any additional information to add or further comments about the survey. 41 participants left a comment, although 14 of them simply said they had no further comments to add.

Five participants noted that there were other locations in which they kept documents, with one identifying a flash drive, and four mentioning network drives or servers. One person mentioned that they use Microsoft Outlook archives as a document repository, while another praised the Outlook interface with its hierarchy and previews, suggesting Windows Explorer could be improved to match.

Four people commented on the survey itself, saying it was interesting, and one person noting that for some questions they would have chosen more than one answer.

Two people mentioned sorting, with one saying they used underscores in file names to force a sort order, and another saying they used numeric codes such as 100, 200, 300 at the beginning of their folder names to enforce a sort order and to make a structure that is easily understandable to others.

One person commented that *"I'm not as tidy as I know I should be!"* while another lamented the difficulty in finding time to organise documents and files properly. One said *"I should probably take some time to organise my files - I would take a mini-course from somebody who has worked out an effective system since I am not about reinventing wheels."*

One person noted that they create short-lived text files on the Desktop for to-do lists or lists of web pages to visit, which are usually deleted within a day. A respondent noted using the Desktop to contain shortcuts to their frequently used folders in My Documents.

One person noted that they *"save EVERYTHING (or just about)"*, including emails. They frequently ran out of storage space, and note that they are sure there are files they haven't accessed in years *"cluttering up my hard disks"*.

Another said they keep everything on their personal laptop which they connect to the University network. This means they can always have their full set of documents with them wherever they are and don't need to worry about having an incomplete set or worrying about synchronisation. The drawback is they need to take responsibility for their own backups.

One respondent observed that they store things different at home to at work. Another noted that as they have just arrived at this University, their system is not yet set up and they have duplicate files.

5.3 FILE SYSTEM SNAPSHOT RESULTS

A total of 78 of the 115 survey participants also completed the file system snapshot. However, five of those only included the default locations of My Documents and Desktop, and didn't add their additional document locations to the snapshot despite indicating in the survey that their primary storage location was a network drive or removable drive. As a consequence, these participants only had a handful of files in the snapshot. These participant's snapshots were removed from this analysis, leaving a total of 73 snapshots for analysis.

A script automatically checked for duplicates in the top level folders. One participant had included their Removable Memory stick twice, and another had included their My Documents Folder twice. In both cases, the duplicate was removed.

The folders were manually scanned to see if there were any anomalies. One participant had a Recycle bin folder on an external drive containing a number of system generated folders and file fragment. This folder was removed from the analysis. Another had included a shared drive that is accessible by the entire Business School. The snapshot program picked up files and folders belonging to other staff members and even other departments. It was impossible to tell from the snapshot data which part of the drive was owned and managed by the participant so the entire drive was excluded from the analysis.

Two participants had included folders that were copies of their entire hard drives, which were created whenever an employee of the Business School was moved to a new computer. These folders include the Windows and Program Files folders containing large numbers of system files and applications. One of these participants had five such folders. These were removed from the analysis.

In the interviews, Brett had a huge number of files and folders in their Visual Studio Projects folder, a folder that is managed by Visual Studio, and includes a large number of folders and files. To better capture just the documents under the participant's direct management, this folder was removed from the analysis in all the interview participants. To ensure consistency, this folder was eliminated from the survey snapshot data also.

A number of applications do sometimes store data in the My Documents folders or its subfolders. However, because of the huge possible range of these application, and the difficulty of deciding between application generated folders and files and human generated ones, no attempt was made to eliminate these except in the case of the Visual Studio Projects folder as discussed above.

The folders that are automatically created by Windows XP (such as My Music, My Pictures) were not removed from the analysis. While some people do not use these, others do, and they are consistently present in all snapshots and therefore don't really introduce any analysis problems.

5.3.1 Overall Size

The mean number of files observed in the document folders was 5,850. However, the number of files ranged from a minimum of only 100 files, to a maximum of 33,902 (standard deviation is 7,605). As **Figure 125** shows, the distribution is significantly right skewed, with a median of only 2,754 and a skewness statistic of 2.26.

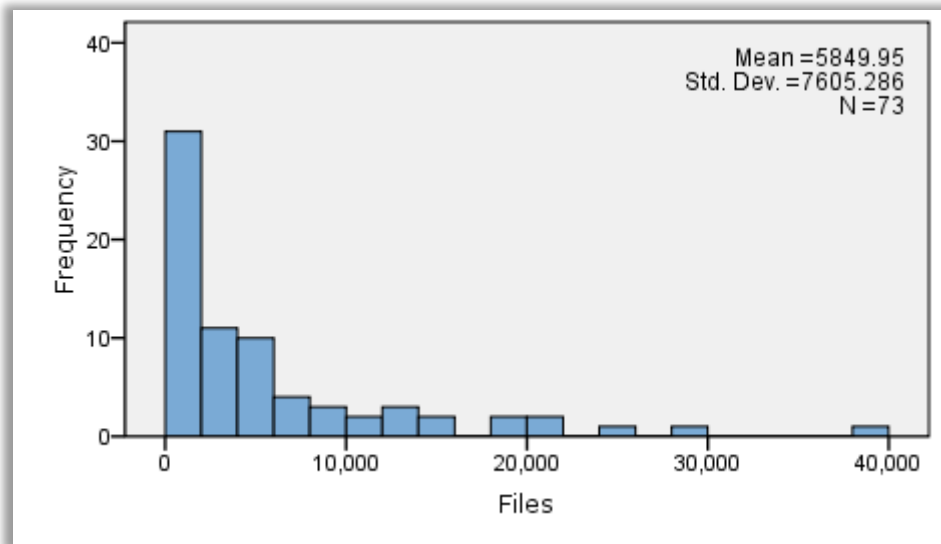


Figure 125: Frequency distribution of total number of files in the snapshot

There is no relationship between the total size of the file system and the participant's satisfaction with their document management practices.

The average number of folders was 628, with a standard deviation of 860. The distribution was also right-skewed (skewness 2.7), with the median number of folders being only 350. The smallest number of folders observed was 11, and the largest was 4,694.

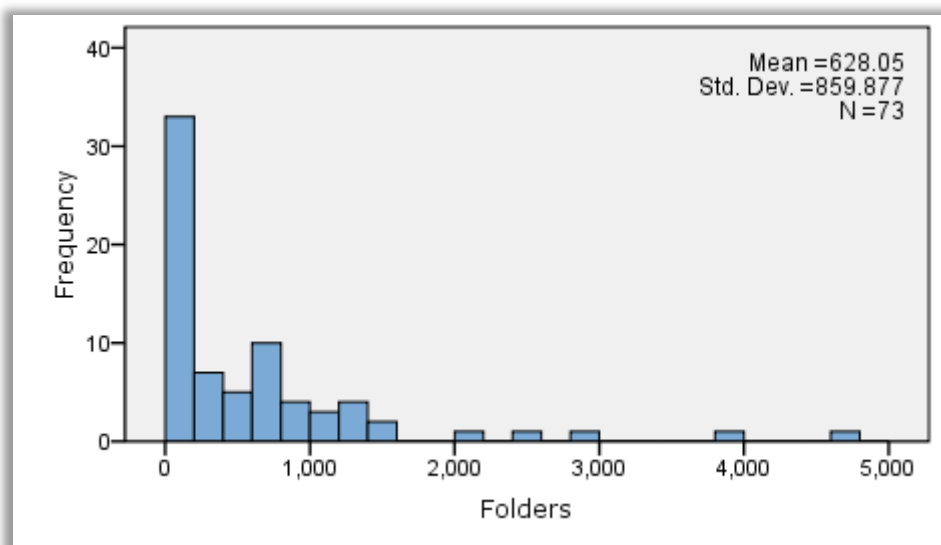


Figure 126: Frequency distribution of total number of folders in the snapshot

The participant with the highest number of folders was not the same person who has the highest number of files. As shown in **Figure 126** there are five participants with over 2,000 folders.

The distribution of files and folders is extremely skewed and non-normal. Transforming the data using a natural logarithm functions results in an approximately normal distribution, therefore this transformed data has been used in all parametric tests that follow.

As would be expected, there is a correlation between the number of files and the number of folders a person has in their file system (correlation coefficient of transformed data is 0.957).

There is no correlation between the number of files or folders a person manages and any of the demographic data collected (age, gender, academic or general staff status, department, length of time they have been working in the same field, or length of time at the University of Auckland).

5.3.2 Locations

The majority of participants (56%) included only 2 locations in the analysis, with a further 33% having 3 locations. One participant had only 1 location, and one participant identified 39 separate locations in the analysis, with the majority of these being folders in the root of the C Drive. Of the 24 snapshots with three locations, in 12 of these cases the locations were the Desktop, My Documents and a personal network drive provided by the university. In a further 11 cases, the locations were the Desktop, My Documents and an additional folder on the local hard drive. The final case had the Desktop, a folder on the local hard drive and a network drive as the three locations.

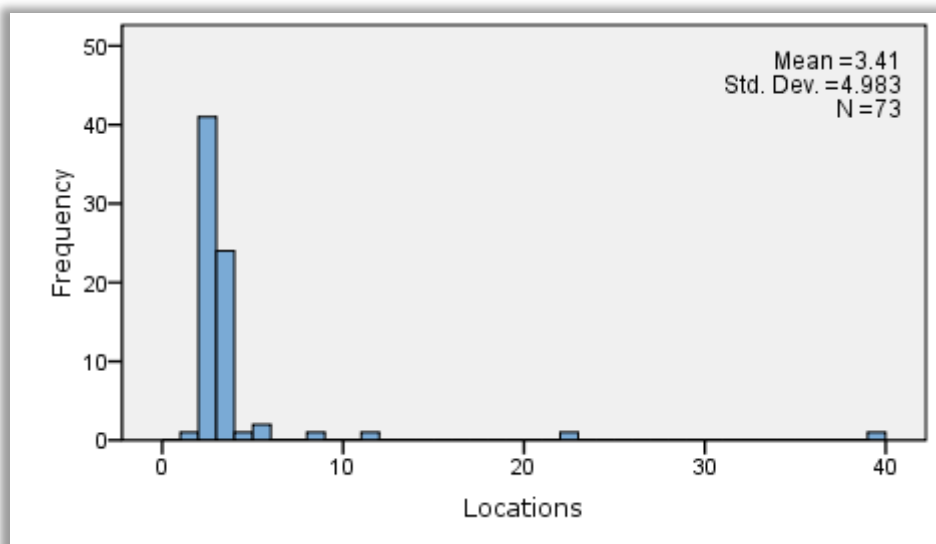


Figure 127: Frequency distribution of the number of top-level locations in each snapshot

It is possible that some of those respondents who only included the two default locations did in fact use additional locations but did not include them in the analysis.

There is no correlation between the number of locations a participant uses and the total number of files and folders they manage.

People were deemed to be using the Desktop for document storage if they had more than five (non-shortcut) documents on the Desktop itself, or if they have more than one subdirectory on the Desktop. By these criteria, it was found that 62% of people are using the Desktop to store documents. However in the survey, only 41% of people reported using the Desktop for document storage.

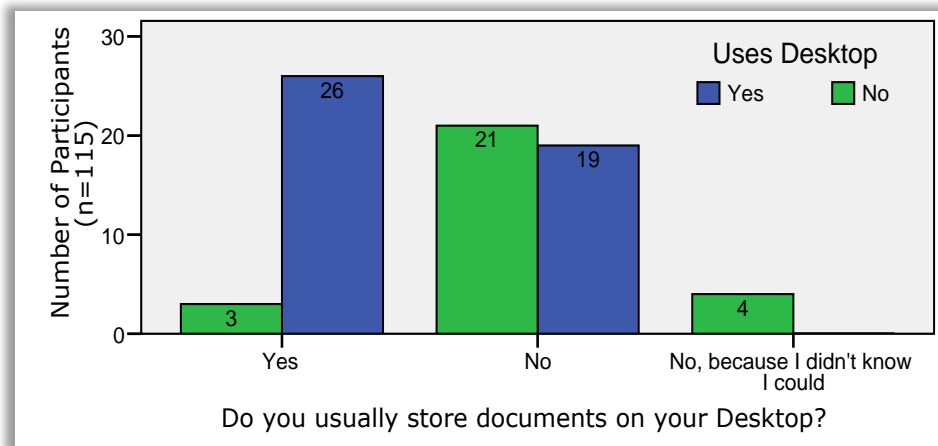


Figure 128: Comparison of actual and reported use of the Desktop

Almost all (90%) of the people who reported that they do store documents on the Desktop were considered Desktop users by these criteria. However, nearly half (47.5%) of the people who stated that they do not use the Desktop to store documents were considered Desktop users. One explanation for this is that although they stated they do not usually store documents there, they may have had some there at the time of the snapshot.

Of these 19 participants, surveying the files on the Desktop showed that 11 of them had a small number of files (typically 5-10) that were generally a mixture of shortcuts to websites, executable files, Microsoft Word documents and PDF documents. The remainder had a larger number of documents that were predominantly Microsoft Word documents, PDF documents, and Excel spreadsheets.

There appeared to be two distinct groups of Desktop users: low users who had fewer than 100 documents on their Desktop (71% of Desktop users) and high users, who had more than 100 documents on their Desktop (making up 29% of Desktop users).

There is a difference in satisfaction with document management between these three groups ($f=4.03$, $\text{sig}=0.02$). Post hoc tests showed that high Desktop users are significantly less satisfied with their document management practices than people who don't use the Desktop ($\text{sig}=0.01$).

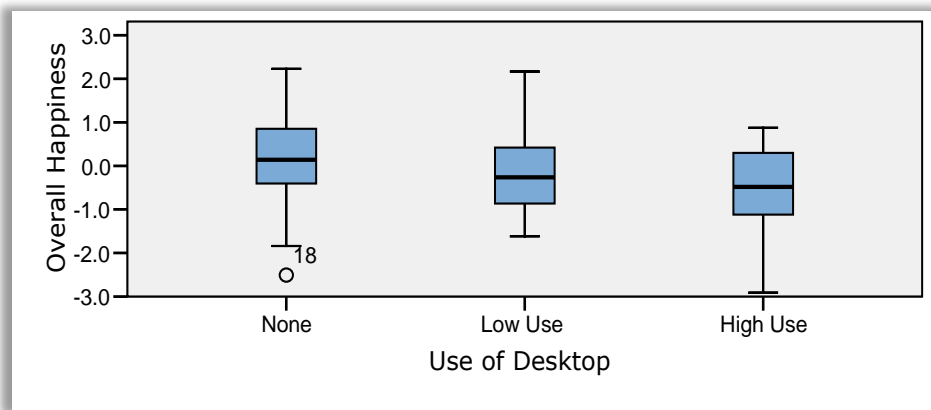


Figure 129: Box plot showing high Desktop users are less satisfied than non-Desktop users

Figure 130 shows how each participant splits their documents between the My Documents folder, the Desktop and Other locations. Around 47% of participants can be considered to use My Documents as their primary document storage location. 44% of participants use some other location as their primary location (most of these are network drives). Only 5.5% use the Desktop as their primary storage location. Another 4% use a combination of two locations. One participant has their documents almost equally split between Desktop and My Documents, and another two have their documents split between My Documents and another location.

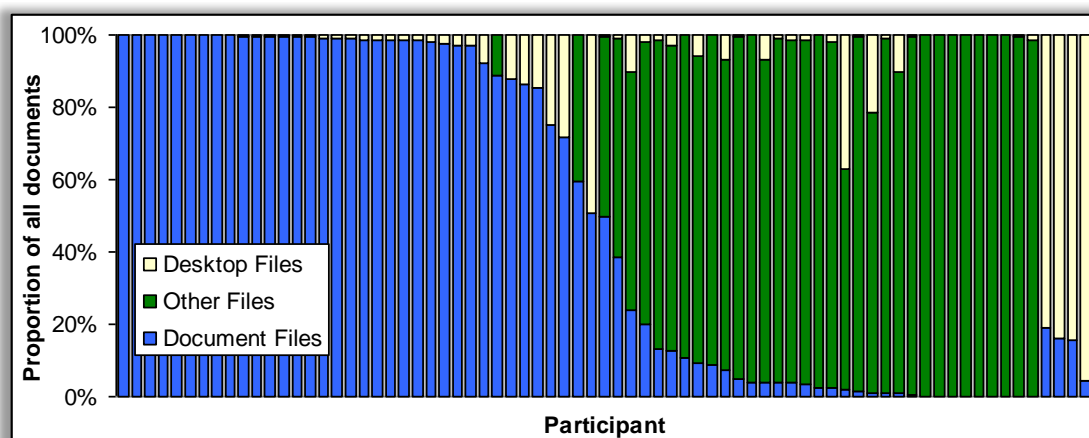


Figure 130: Allocation of files between Desktop, My Documents and other locations

There is a difference in satisfaction depending on the participant's primary storage location ($f=2.46$, $\text{sig}=0.071$). Post hoc tests (LSD) indicate that the participants who use a primary storage location other than My Documents or the Desktop are happier with their document management than people who use My Documents ($\text{sig}=0.011$).

5.3.3 Depth

There is significant variation in the depths of folder structures. The shallowest structure was only 1 level deep, while the deepest was 18 levels deep. The mean maximum depth was 6.8 levels, with a standard deviation of 3.1.

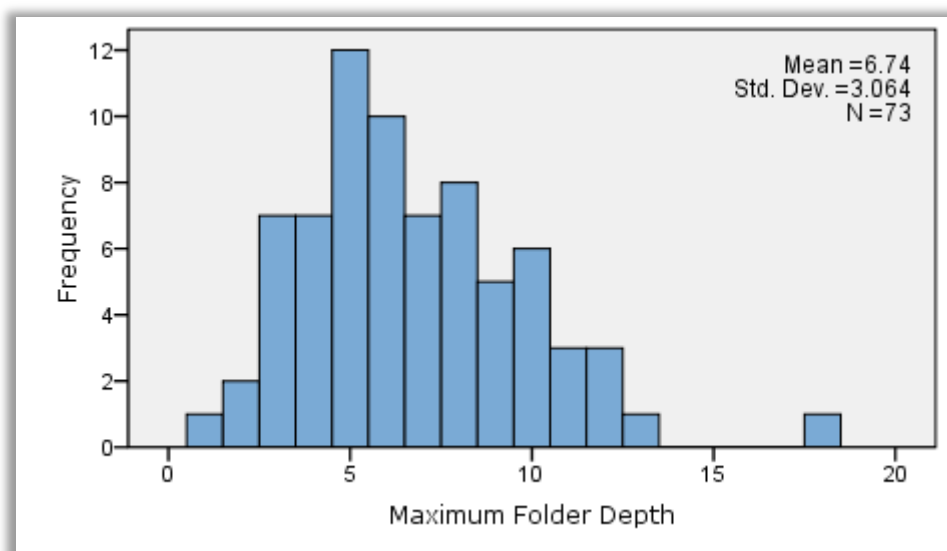


Figure 131: Bar graph showing distribution of maximum folder depth

Since file systems structures are unlikely to be uniformly deep, the mean depth might be a better metric than the maximum depth. There are two ways this can be calculated. The first method simply averages the depth of every folder in the structure. The second method only considers the leaf folders – folders that do not have any subfolders. In practice, there is very little difference between the two methods, with the two being very highly correlated ($r = 0.99$).

For mean leaf depth, the highest value recorded was 9.35 levels, with an average of 3.4 levels and a standard deviation of 1.2. This participant with the highest average value was the same participant who had an 18 level deep structure. The average depth of non-leaf folders across all participants was 3.6, with a standard deviation of 1.2. The highest non-leaf average depth was 9.65.

The maximum depth is positively correlated with both the total number of files ($r = 0.65$) and the total number of folders in the file system ($r = 0.68$).

5.3.4 Bushiness

On average 74% of folders did not contain any subfolders at all, only (possibly) files. These are considered leaf folders. The lowest proportion of leaf folders observed in a file system was 50% and the highest was 90%.

The interior (non-leaf) folders by definition must contain at least one subfolder. The mean number of subfolders per folder was 4.1, with a standard deviation of 1.3. The highest average observed was 9.5, and the lowest was 1.8 subfolders per folder.

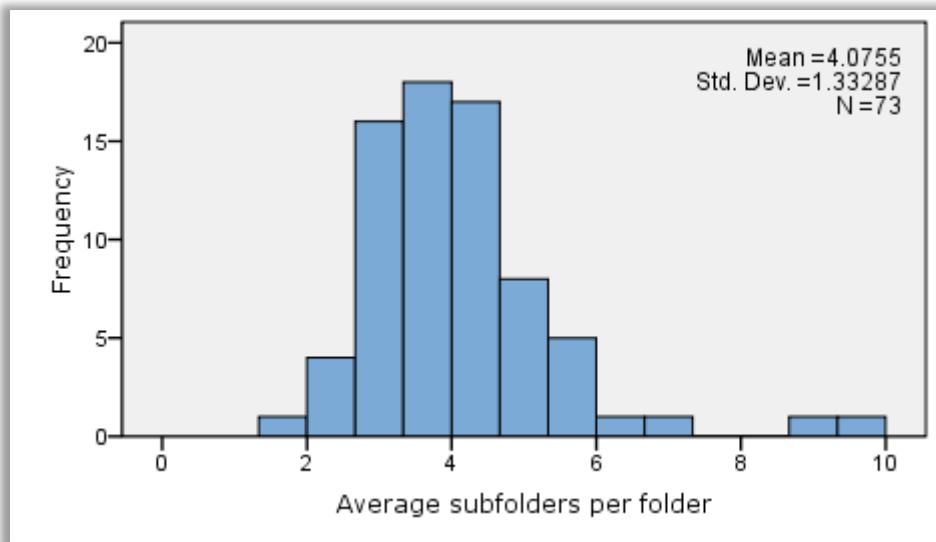


Figure 132: Histogram showing distribution of mean number of subfolders per folder

There is no correlation between the average number of subfolders per folder and the total number of files and folders in the file system. There is also no significant correlation between the average depth of the tree and the average number of subfolders. Since depth does vary with the total size, this would imply that the bushiness of the tree varies independently of these factors.

There is, however, a significant correlation between the maximum number of subfolders per folder and the total number of files ($r = 0.55$) and folders ($r = 0.62$) in the file system.

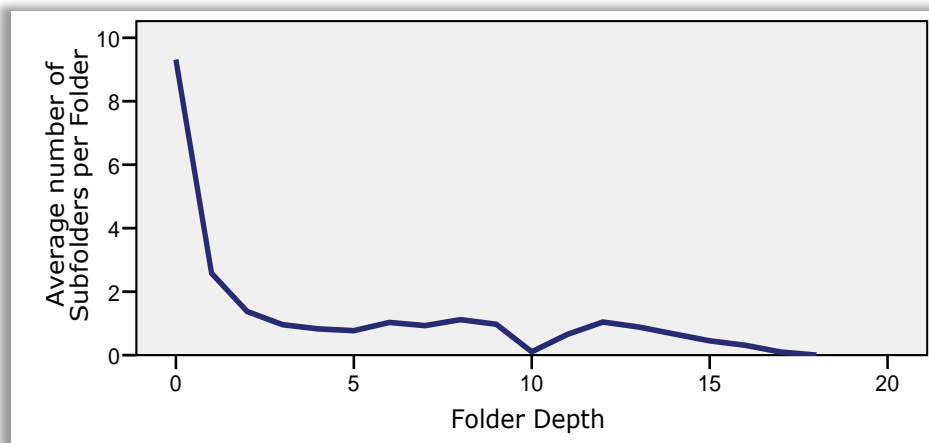


Figure 133: Line graph showing how the average number of subfolders in a folder varies with the depth of the folder

The most significant feature of **Figure 133** is the high average number of subfolders at the root of the tree, which sharply drops off by two or three folders down. The dip at 10 is just an artefact of lack of data; since only 8 participants have a file system more than 10 levels deep.

This relationship between depth and number of subfolders can also be compared across the three different document collections: Desktop, My Documents and other locations.

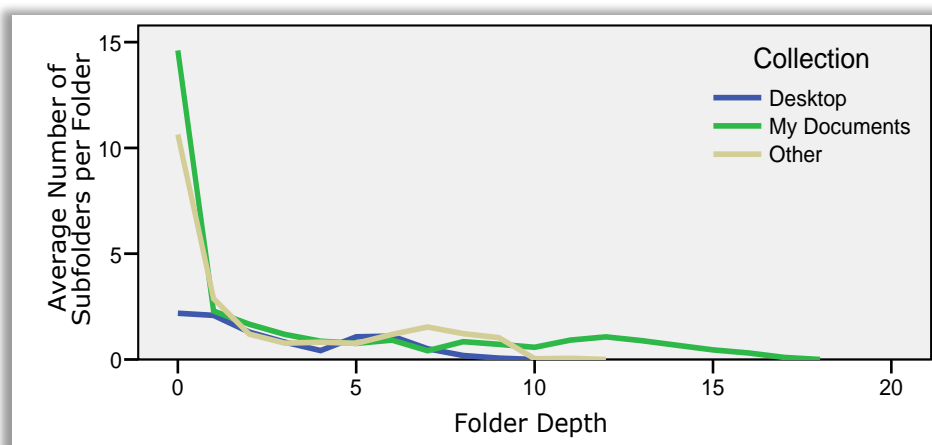


Figure 134: How the average number of subfolders in a folder varies with the depth of the folder across each of the three collections

As **Figure 134** shows, the Desktop differs from other documents in having a much smaller average number of subfolders on it.

Another metric of bushiness is known as the branching factor. If you took all the folders in the document structure and redistributed them into a completely uniform tree with all branches having the maximum height and all folders having the same number of subfolders, then the branching factor tells you how many subfolders each folder will have. For the purposes of calculating this branching factor, the top level locations were considered to be level 1, with a 'virtual' level 0 considered to be the single root of the user's document space.

The average branching factor was 1.93. This ranged from 1.27 to 2.97 and had a standard deviation of 0.34.

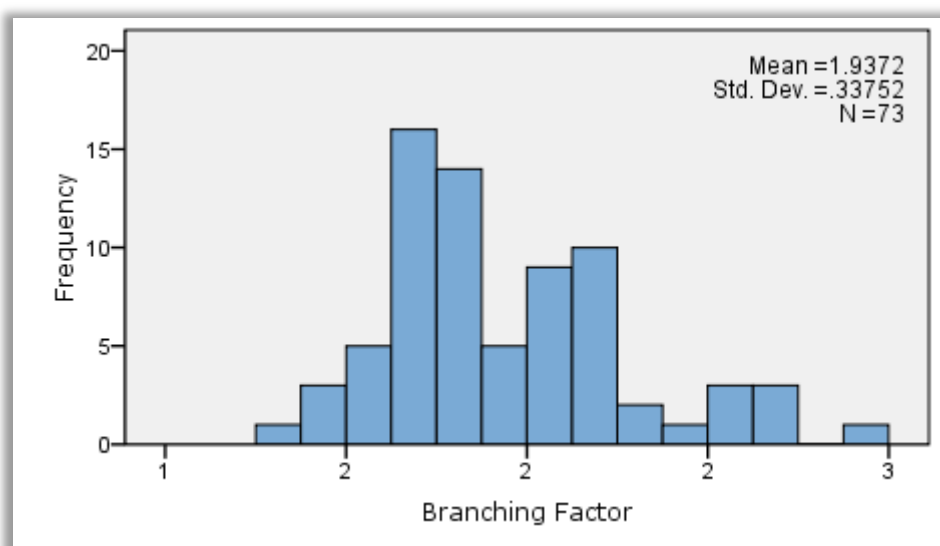


Figure 135: Histogram showing distribution of branching factor

As would be expected for two measures of bushiness, the branching factor and the average number of subfolders are highly correlated ($r = 0.5$).

In common with the average subfolders metric of bushiness, there is no correlation between the branching factor and the total number of files and folders in the file system.

There is a small but significant negative correlation between the branching factor and the average depth of the tree ($r = -0.36$), indicating that wider trees tend to be shallower. There is also a positive correlation between the branching factor and the number of top level locations ($r = 0.41$). This is expected, since the locations essentially represent the top level of tree branching.

Branching factor has a weak correlation with satisfaction. Although it is statistically significant, the correlation coefficient is only .334 ($\text{sig}=0.004$).

One of the key differences between the branching factor and the average subfolders is that branching factor assumes a perfectly even tree, whereas the average number of subfolders is affected by the tree's unevenness. This aspect of the tree structure will be examined separately in the Balance section below.

To further analyse the bushiness, the following graph shows that across the entire range of participants, 68.4% of folders are leaf folders, having no subfolders. 15.6% of folders have a single subfolder, 6.5% have two or more and then the proportions drop off from there. A very small percentage of folders (0.4%) contain more than 20 subfolders, and even fewer contain more than 100 (only 7 folders across all participants, a negligible amount).

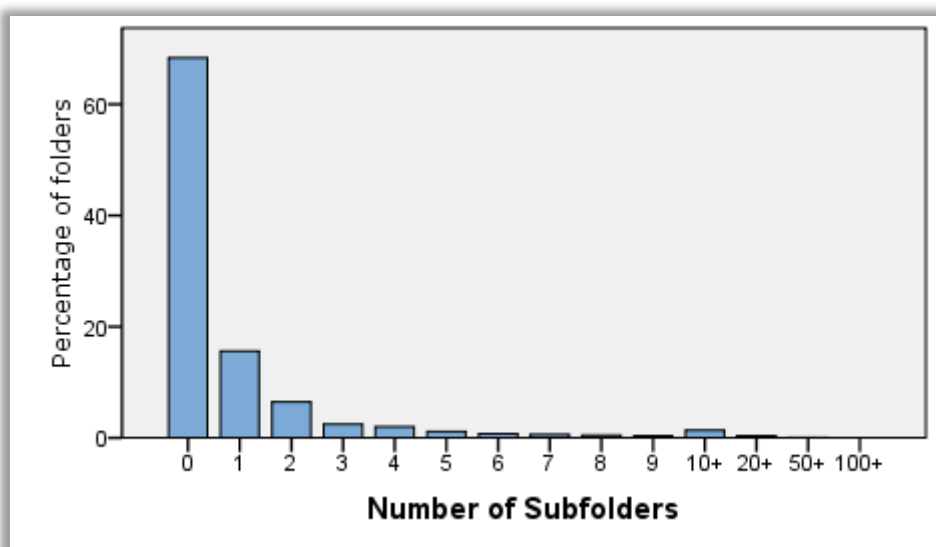


Figure 136: Bar graph showing the percentage of folders containing different numbers of subfolders

Rather than treat all folders equally, we can go into more detail and look at how the proportion varies for each participant. Although **Figure 137** below is very cluttered with lines for each participant, you can see that there is a notable fan shape, indicating that those participants with the largest proportion of leaf folders (highest percentage at level 0) tend to have the fewest number of folders containing subfolders (lowest percentage at level 1 and above). This implies they have fairly shallow

structures. By contrast, there are participants who have around 50-60% of the folders as leaves, and the rest contain subfolders, giving them a deeper tree structure.

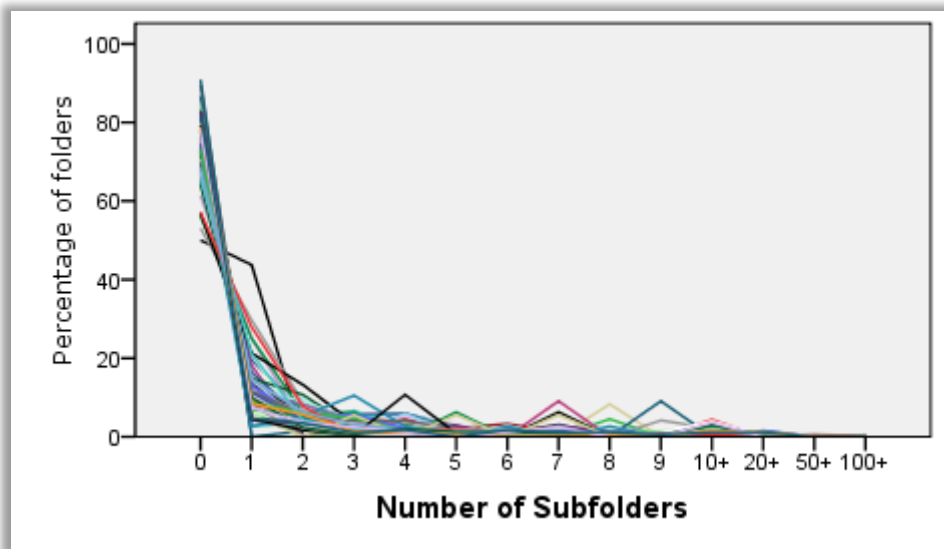


Figure 137: Graph showing the percentage of folders containing different numbers of subfolders for each participant

We can see the pattern more clearly if we zoom in to **Figure 138** below to show just the folders containing none, one or two subfolders. One extreme participant has almost 100% of their folders as leaves – they have just a single top level folder containing all their other folders, none of which have any further subfolders. This participant is an extreme example, having the second smallest number of files of all participants, only 101 files.

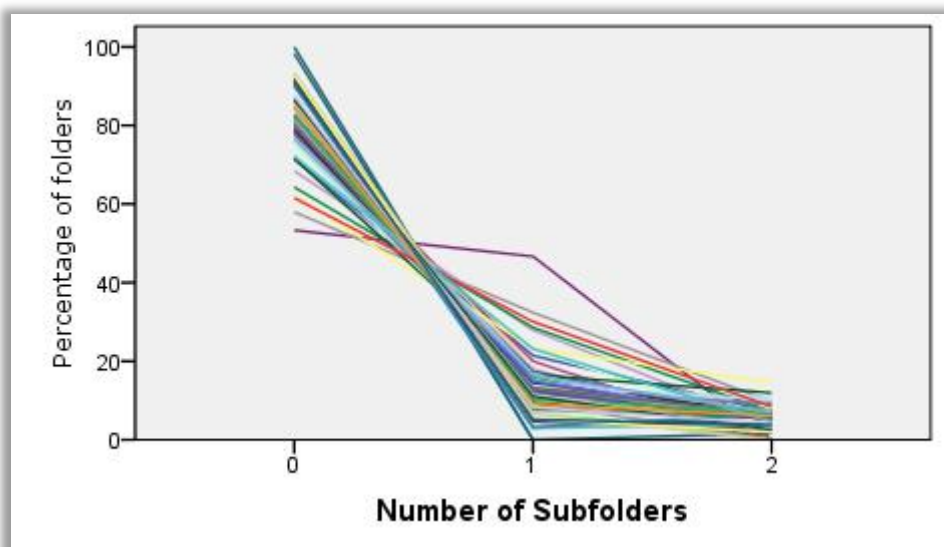


Figure 138: Graph showing the percentage of folders containing zero, one or two subfolders for each participant

One participant is a notable outlier, having only 55% of their folders being leaf folders, and 50% of their folders containing one subfolder. This participant actually barely uses folders at all, keeping almost

all of their 1028 files directly in the My Documents folder. There are only a few folders in this folder, mostly created by Windows XP or various applications.

5.3.5 Leafiness

The average number of files per folder across all file systems was 11.1. The highest number of files observed in a single folder was 1168. All of the files in this folder were JPEG images.

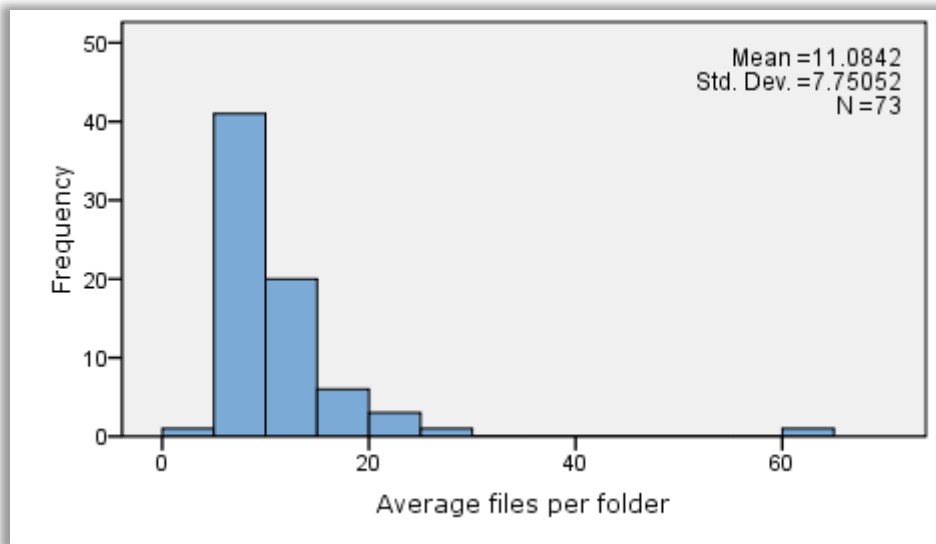


Figure 139: Histogram showing distribution of mean number of files per folder

The least leafy file system had an average of 4.5 files per folder, and the leafiest averaged 64.3 files per folder. However, this was a significant outlier, with the second leafiest file system averaging under 30 files per folder.

It seems natural to expect that trees that are bushier (wider) would also be shallower, and deeper tree structures would be narrower. However no correlation was found between the average number of files per folder and the overall number of files, nor with the average depth or bushiness of the document structure. However, similar to bushiness, there is a significant correlation between the maximum number of files per folder and total number of files ($r = 0.56$).

Rather than comparing the average values for each file system, we can also compare using individual folders. As shown in **Figure 140**, the average number of files per folder is highest at the top levels of the tree, with the main location folders averaging over 30 file each. After that it drops off sharply, with the first few levels of subfolders averaging around 10-12 files each. It is fairly constant at levels 1 to 5 of the tree and then tapers off. Note that the average file system has a maximum depth of 6.8 levels, so values much beyond that are not really representative due to lack of data.

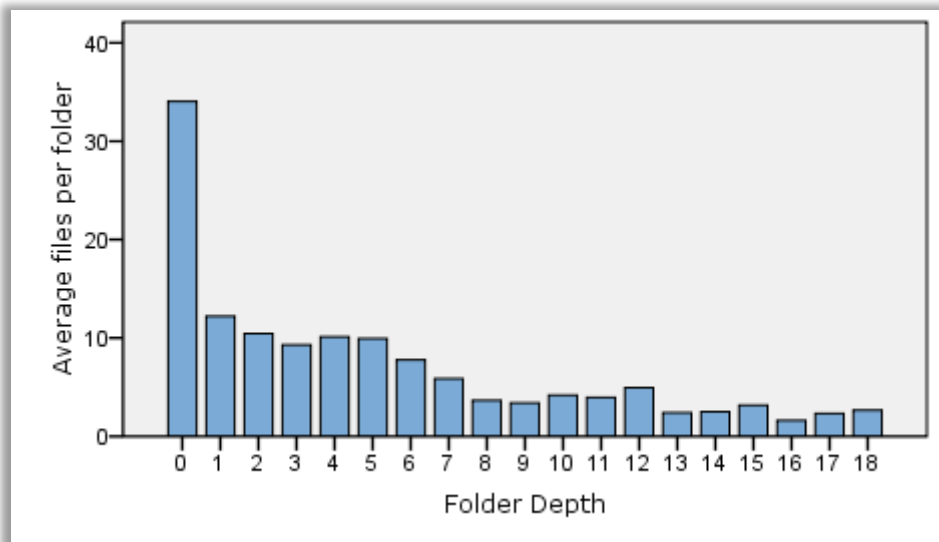


Figure 140: How the average number of files in a folder varies with the depth of the folder

While this shows us the per folder statistics at each level, it doesn't give a picture of just how many files are at each level of the tree. Because of the much higher number of folders at intermediate levels of the tree, the total number of files stored there is high. **Figure 141** below shows that most files live at depths of 2 to 5 in the tree.

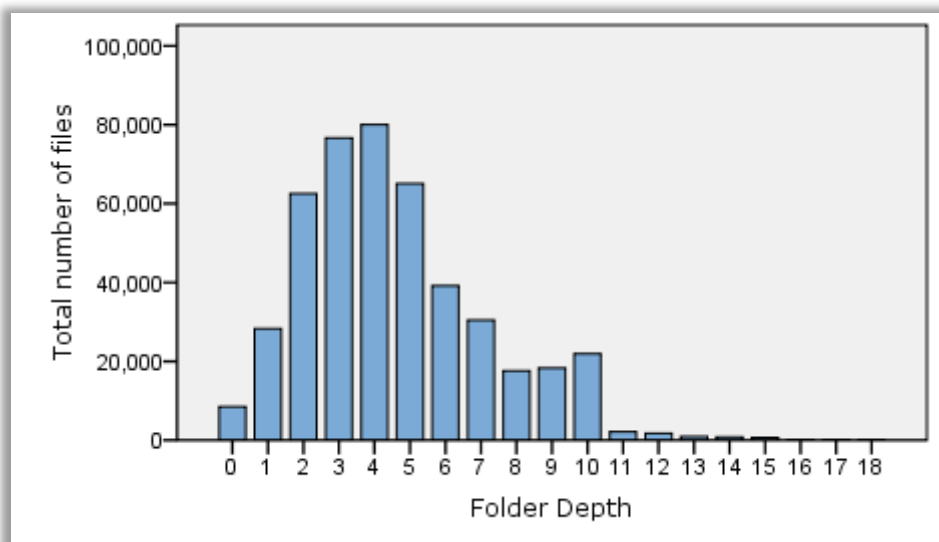


Figure 141: Bar graph showing total number of files stored at each folder depth

5.3.6 Balance

To assess how even the distribution is across the tree, the standard deviation of the number of subfolders was used as a measure of balance. The lower the standard deviation, the more evenly balanced the tree. The average balance was 5.6, with a standard deviation of 3.5. While this indicates that most trees are fairly balanced, there is one significant outlier having a standard deviation of 27.9. This person has a relatively small file system of 628 folders in total. While most of their folders only

have one or two subfolders, they also have folders that contain 105, 142 and 151 subfolders, giving them an extremely unbalanced tree.

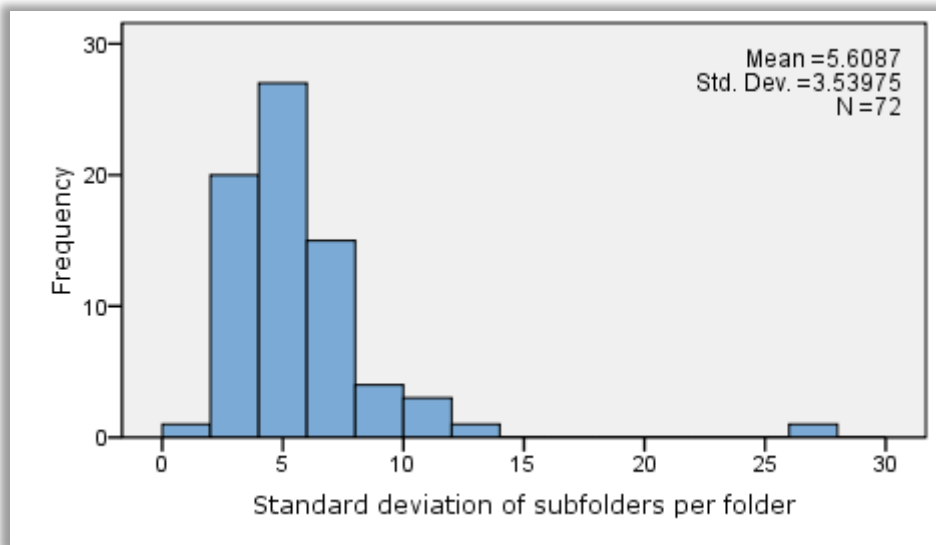


Figure 142: Histogram showing the distribution of the standard deviation of the number of subfolders

There is no correlation between the balance of the tree and the overall size or the number of top level locations. Nor was there any relationship with the depth of the tree.

There is a statistically significant correlation between the balance of the tree and the bushiness, using both the average subfolder metric ($r = 0.73$) and the branching factor metric ($r = 0.42$). This would indicate that trees that are wider on average also tend to be less evenly balanced than narrower trees.

In addition to assessing how balanced the folder structure is, we can also examine how evenly distributed files are throughout the tree. The standard deviation of the number of files in a folder was used as a measure of balance. The average file balance was 23.4, with a standard deviation of 29.5. As with the folders, this was also considerably right skewed due to one outlier. This participant has 966 out of their total 1028 files in the My Documents folder. They have not created any folders to structure these documents, and do not appear to have made use of any of the system created folders to organise these documents.

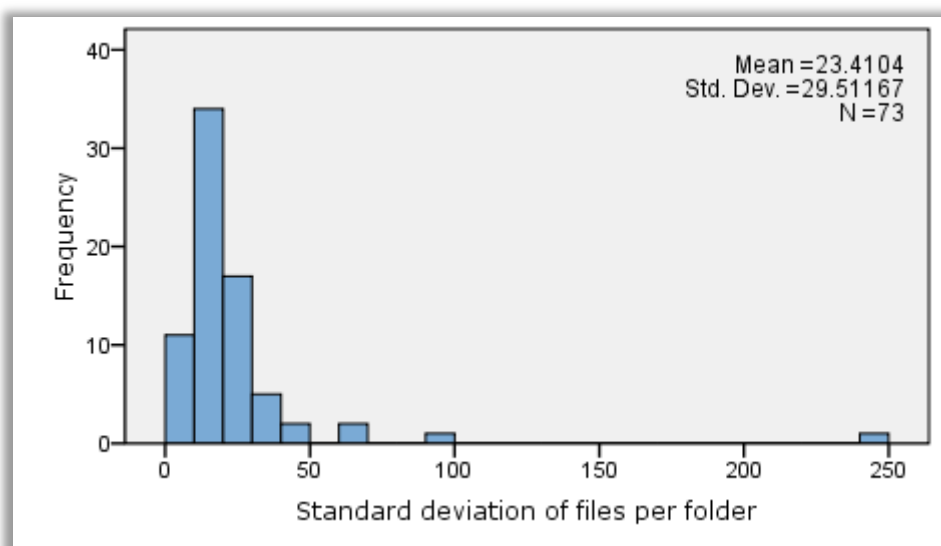


Figure 143: Histogram showing the distribution of the standard deviation of the number of files

There is no correlation between how evenly the files are distributed and the balance of the tree structure itself. There is also no correlation between the file spread and the overall size or depth of the document structure

There is a statistically significant correlation between the file balance of the tree and the leafiness ($r = 0.93$). This would indicate that trees that are wider on average also tend to be less evenly balanced than narrower trees.

5.3.7 Emptiness

Only three file systems did not contain any empty folders. The highest number of empty folders was 610, as shown in **Figure 144**. Of these, 280 appear to be system generated folders with extremely similar names. The rest do not appear to follow any particular pattern, and occur at all levels of the folder structure. The participant with this number of empty folders has the highest total number of folders (4,694) and the second highest number of files (29,998).

The mean number of empty folders was 37.8 (s.d. 83.0). The distribution of the number of empty folder was extremely right-skewed (skewness 5.4), with a median number of empty folders being 13.

Since the number of empty folders can be expected to increase with the size of the file system, the proportion of empty folders is perhaps more important. The emptiest file system had 47.6% of the folders being empty.

In this case, the participant has a total of 145 folders, and they have quite a few empty subdirectories of My Documents which appear to have been created by either Windows itself, or various imaging and multimedia applications. Some of these had extensive subdirectory structures that were largely (or completely) empty.

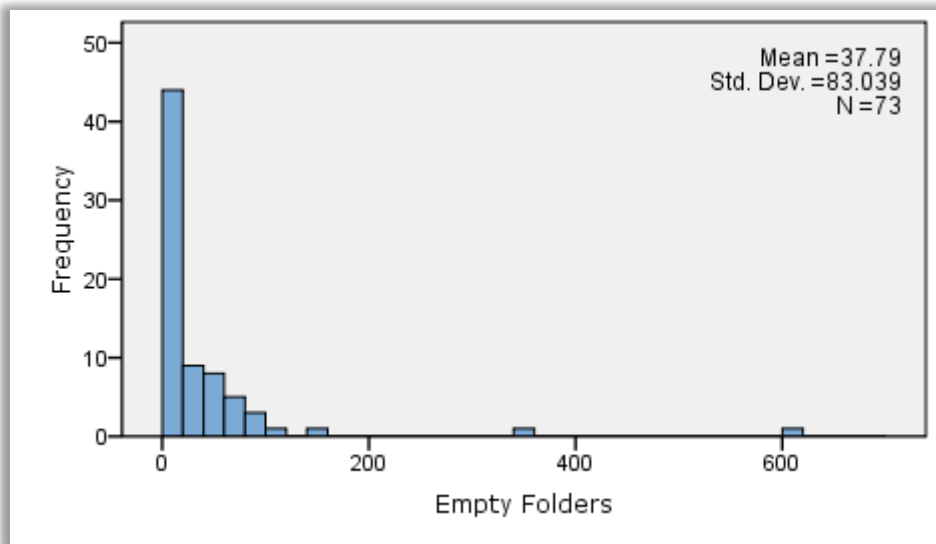


Figure 144: Histogram showing number of empty folders

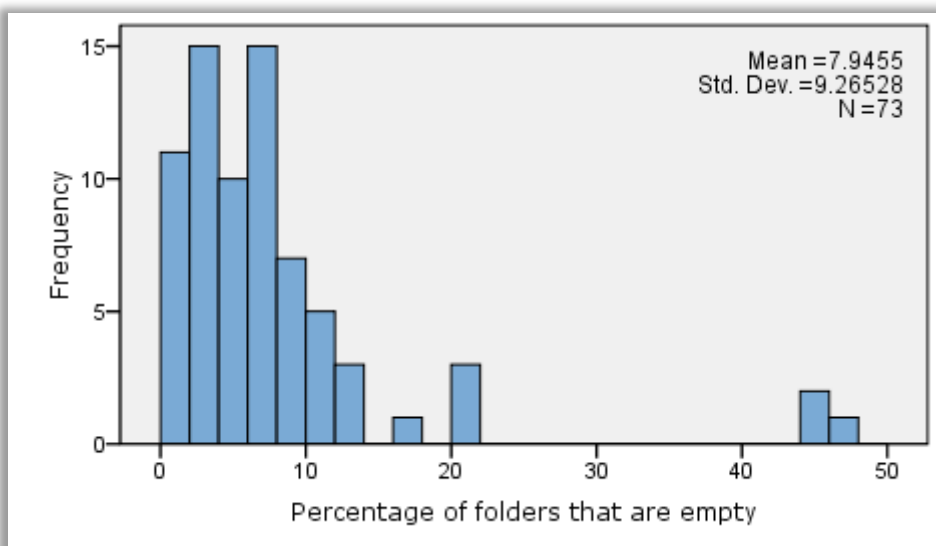


Figure 145: Histogram showing percentage of folders that are empty

Most file systems had only a small proportion of empty folders, with the mean proportion of empty folders being 7.9%.

The proportion of empty folders is not related to the overall size of the document structure, or to the bushiness or leafiness of the structure. There is a weak negative correlation between the proportion of empty folders and the average depth of the file system ($r = -.27$, $p = 0.02$). This indicates that shallower structures are more likely to have empty folders than deeper ones.

5.3.8 Duplication

Duplication was measured by calculating the proportion of non-unique file names in the file system. A file or folder is considered to be a duplicate if another file or folder exists with the same name

anywhere else in the file system. It is entirely possible that the contents of the file or folder may be completely different. This duplication may be accidental or deliberate.

The mean level of file duplication was 21.8%. This means that on average, 21.8% of the documents in the file system have the same name as another file. The amount of file name duplication ranged from 0.4% to 60.4%.

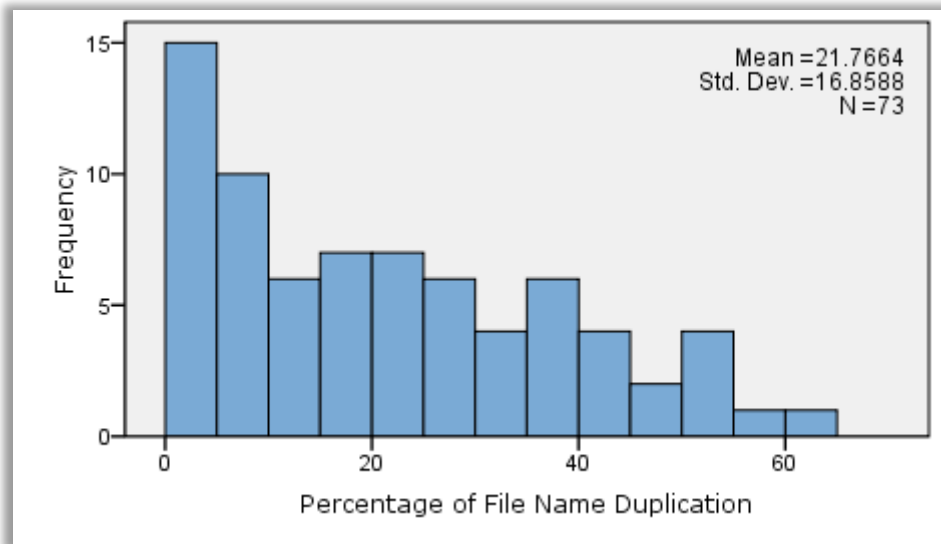


Figure 146: Histogram showing distribution of proportion of file name duplication

The level of folder name duplication was slightly higher, with a mean of 23.5%, and ranging from 0 to 73.4%.

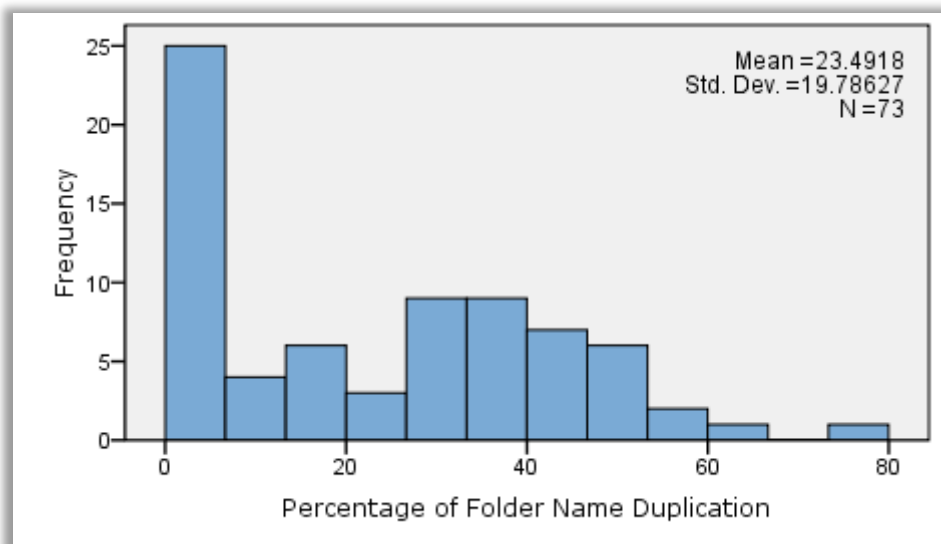


Figure 147: Histogram showing distribution of folder name duplication

There is a significant correlation between the level of folder and file name duplication ($r = 0.79$). One explanation for this might be that entire folders and their contents are being duplicated together.

The participants are fairly evenly split in whether the duplication is higher in files or folders, with 48% having higher folder duplication than file duplication. On average, the amount of folder duplication is about 2% higher than the amount of file duplication. However, the range is quite wide, with the participant at one extreme having 34% more file duplication than folder duplication, to the participant at the other end of the spectrum having 33% more folder duplication than file duplication.

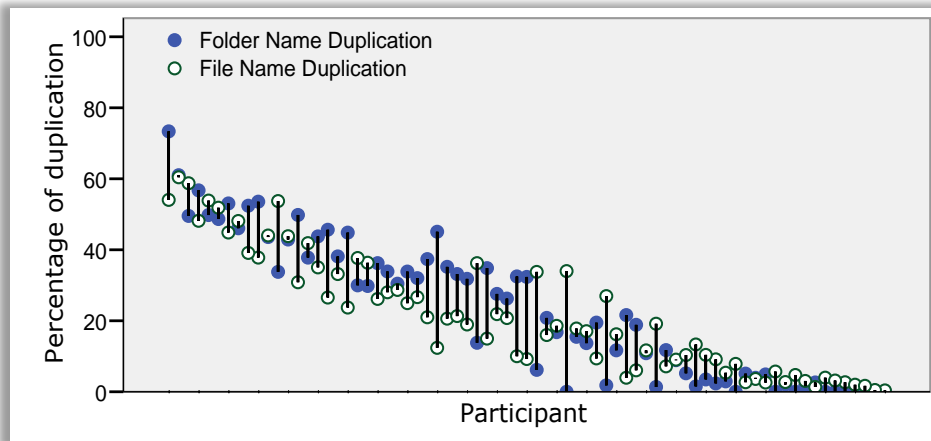


Figure 148: Graph showing relationship between folder name duplication and file name duplication

The amount of duplication is related to the overall size of the file system. The level of file name duplication is correlated to the total number of files ($r = 0.61$) and the level of folder name duplication is correlated with the total number of folders ($r = 0.65$). Therefore, the more folders a person has, the more they are likely to have empty folders.

There is no significant correlation between the levels of file and folder duplication and the leafiness or bushiness of the file system. However, there is a significant correlation between the average depth of the file system and the level of file name duplication ($r = 0.59$) and folder name duplication ($r = 0.71$). This could suggest that people with deep file systems are more likely to have repeating groups of folders and files (which perhaps are differentiated with a high level folder name).

The level of duplication can also be examined within specific locations as well as across the file system as a whole. On the Desktop, the average level of folder duplication is only 2%, and the level of file duplication is 9.7%. Within the My Documents folder, the folder duplication is 14% but the file name duplication is only 2.2%. Within other locations (network drives, flash memory, other C: drive folders), the folder name duplication was 10.5%, and file name duplication is 17.6%.

All of these within-location levels of duplication are lower than the average level of duplication across the whole file system, indicating that files and folders are being duplicated across locations. This can be done for a number of reasons, including portability and backup purposes.

A limitation of this analysis is that duplication is determined solely by having the same folder or file name. There are legitimate reasons why two folders that are not the same may have the same name, and these would be erroneously counted as being duplicates when in fact they are not.

5.3.9 Shortcuts

The Windows file system is a strictly hierarchical tree, which means that each file or folder can only exist at one location in the hierarchy. To facilitate more flexible navigation, the user can create shortcuts to files or folders from anywhere in their tree structure. This allows the tree to become a network, and allows more associations between items to be modelled. Because many applications create shortcuts on the Desktop itself, shortcuts here do not necessarily indicate the user is trying to create a more flexible document structure. Shortcuts within My Documents (or other document locations) or within subfolders of the Desktop are better indicators of attempts to use a more flexible structure. Theoretically, the ability to create shortcuts to a file or folder from other locations means that the file system is actually a graph rather than a tree. Knowing how frequently this option is used might point to fruitful changes in the conceptual structure of the file system.

Shortcuts can either be shortcuts to documents or to applications. The target of a shortcut (.lnk) file is actually encoded inside the file itself. Since the snapshot did not collect the contents of any files, this information could not be collected in the snapshot. Some applications and the operating system occasionally create shortcuts in the document folders (e.g. Links to sample sounds and picture files in My Documents). Therefore, a small number of shortcuts probably indicate that these have not been explicitly and consciously created by the participant.

All but three participants had at least one shortcut on their Desktop. 51 participants (69.9%) had between 1 and 10 shortcuts, a further 14 (19%) had between 11 and 20, with the highest number of Desktop shortcuts being 112. **Figure 149** shows the distribution of the number of shortcuts that each participant has in their system.

Inside the My Documents folders, 14 participants (19%) had no shortcuts at all. The majority (37 participants, 50.7%) had either 1 or 2 shortcuts. A further 15 people had between 2 and 20 shortcuts, 3 participants had between 20 and 50, and 4 had over 50 shortcuts, with the highest being 210. In other folders, the majority (55 participants, 75%) had no shortcuts. 16 people had between 1 and 5 shortcuts, with 1 participant having 82 shortcuts and 1 having 455 shortcuts.

There is a weak correlation between the total number of folders in the file system and the number of shortcuts in the My Documents folder tree ($r = 0.25$, $p = 0.03$). There is no correlation between the total number of files and the number of shortcuts. There is also a small positive correlation between the number of My Documents shortcuts and the average depth of the tree ($r = 0.31$).

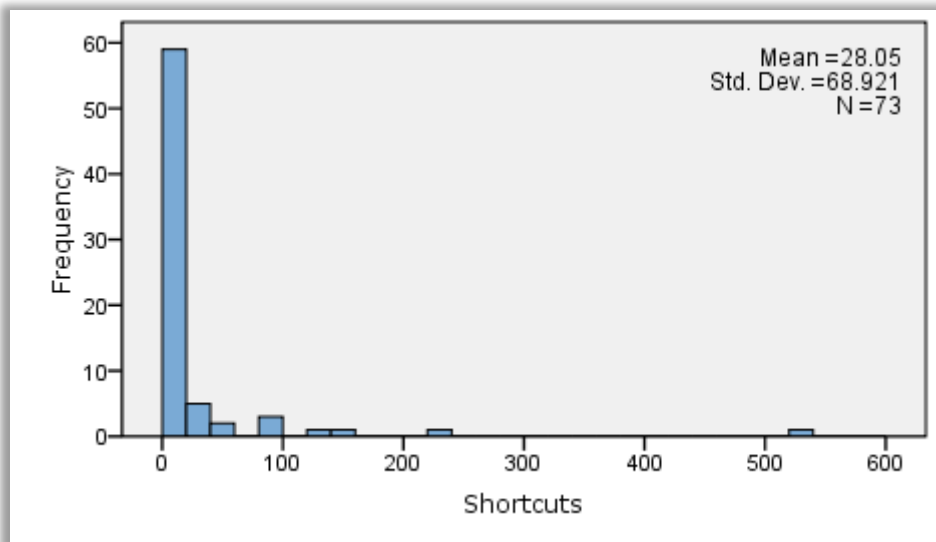


Figure 149: Histogram showing the total number of shortcuts

There is no correlation between the number of shortcuts and the bushiness and leafiness of the structure, or with the proportion of empty files or level of duplication.

5.3.10 File Types

File types were classified according to the extension given to the file. These extensions are typically three or four letters and many document formats have standard extensions.

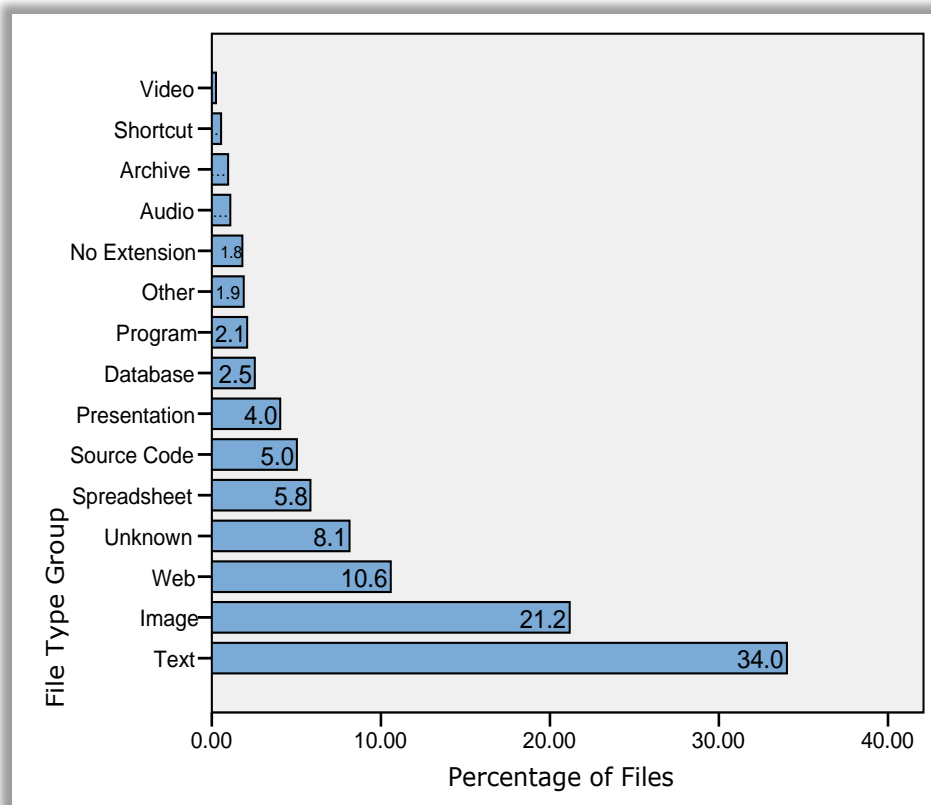


Figure 150: Bar graph showing percentage of files by file type group

For instance, Microsoft Word documents usually have a .doc extension, and Adobe Portable Document Format documents have a .pdf extension. Not all file extensions can be used to unambiguously determine the type of document. For instance, .sav and .log extensions are used by many applications.

Only file extensions used for more than 2 files and in more than two different snapshots were included in this analysis. This provided a total of 623 separate file extensions. In addition, 1.8% of all files (7,582 files) had no file extension at all. 8.1% of files had an extension that could not be unambiguously classified into a particular file class.

90.1% of files had extensions which could be grouped into one of 13 file classes:

- Text. This was the largest category, taking up 34% of all files. Within this category, Microsoft Word and Adobe Portable Document Format were the most common types of documents.
- Image. Images made up just over 21% of the total document collection, with GIF and JPEG being the most common formats.
- Web. Just over 10% of the documents were HTML files.
- Spreadsheet. The 5.8% of the files that were spreadsheets were largely Microsoft Excel Files.
- Source Code. 5.0% of the files were source code files. Among the most prevalent were C#, Java, Pascal, PHP and JavaScript.
- Presentation. Microsoft PowerPoint files made up the 4% of all files that were presentation files.
- Database. The database files primarily consisted of Microsoft Access files, and some Oracle data files.
- Program. Program files such as executables and batch files made up 2.1% of all files.
- Other. The other category consists of file extensions which can be assigned to groups, but which appear in such low frequencies that they have been grouped together for analysis. These include temporary files, backup files, configuration files, xml files and Microsoft Project files.
- Audio. The primary audio formats were MP3 and WMA (Windows Media Audio).
- Archive. Archive files primarily consist of ZIP files, with a much smaller number of other archive formats.
- Shortcut. Shortcuts included links to local and remote files, and also shortcuts to websites.
- Video. The video files were a mixture of AVI, MPG, WMV and MOV files.

Not all file types were present in every participant's file system. As **Figure 151** shows below, all participants had some text type files, and most had spreadsheets, images, presentations, shortcuts, other files and unknown files. Video and Audio files were present in the smallest number of snapshots.

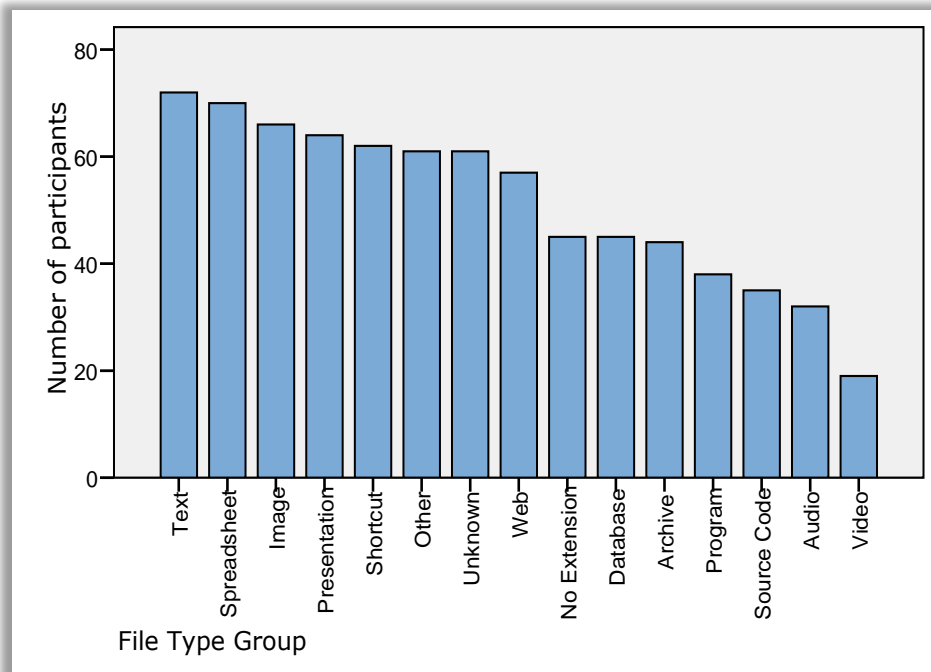


Figure 151: Bar graph showing number of participants who have files of each file type group

Another aspect of file type is whether a particular type tends to occur more frequently as the majority occupant of its folder, or whether it tends to occupy only a lower proportion of its folder, cohabiting with other file types. **Figure 152** shows the average proportion of a folder that each file type occupies.

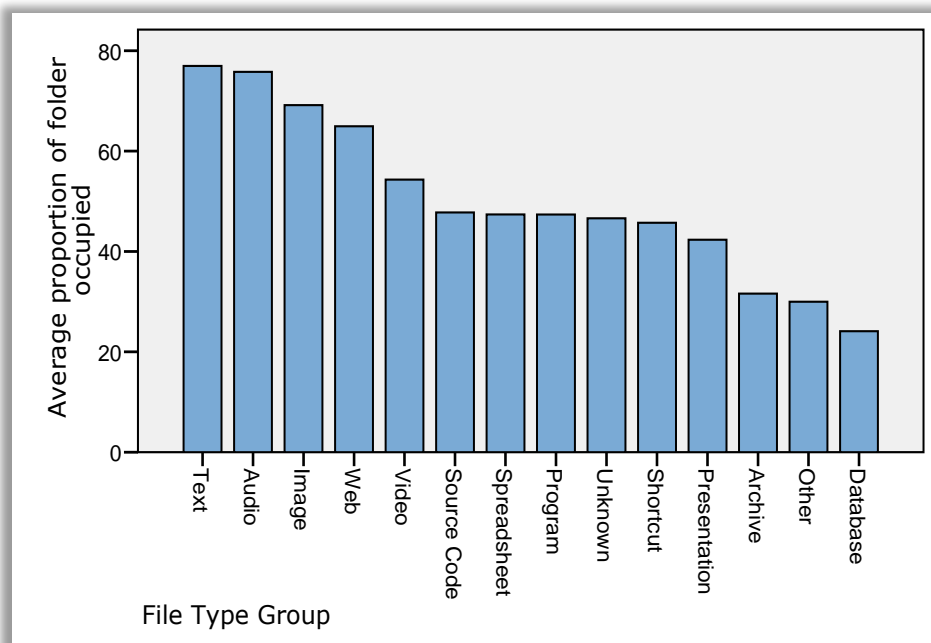


Figure 152: Bar graph showing average folder occupancy by file type group

Text files tend to occupy the majority (77%) of the folders in which they are located. Audio files, Image files and Web folders also have a tendency to be the majority occupants of the folders they inhabit.

5.3.11 File Names

The Windows operating system allows file names to be up to 255 characters long. Typically each file name has an extension that can be used to identify the type of file and which determines which applications will be launched to edit or view the file. By default, Windows hides the file extensions; so many users never actually see them. For the purposes of this analysis, the extension was stripped off, and just the file identifier was used. The file name must be unique within a folder, but files in different folders can have the same name.

In early operating systems, file names were limited to 8 characters for a file identifier followed by a period and then 3 characters for a file type identifier (extension). Window 95 introduced 255-character length file names to the Windows platform. **Figure 153** shows the distribution of file name lengths across all the snapshots.

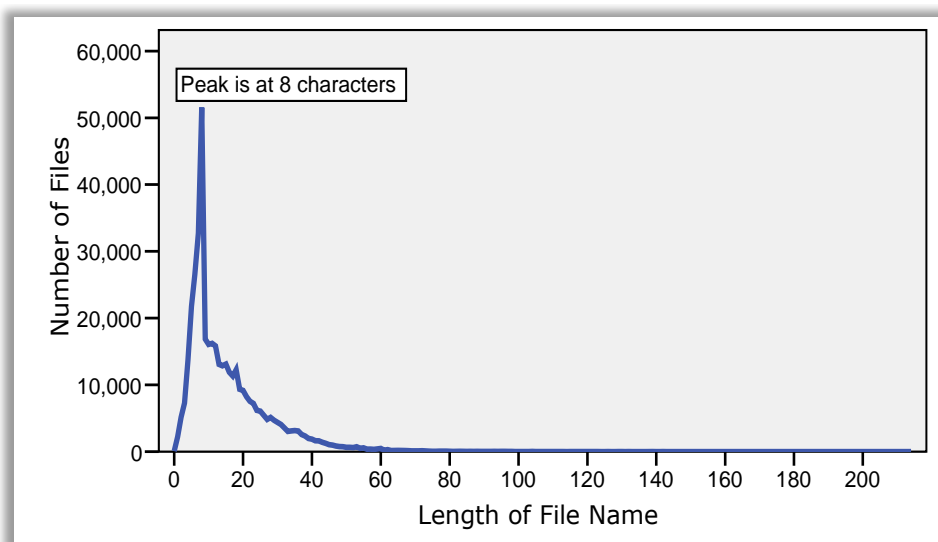


Figure 153: Line graph showing distribution of file name length

The most common file length is still 8 characters. However, you can see that there is a long tail of file names with very long names, with the longest being 214 characters long. There appear to be 69 files with no filename. These do actually have a name; however the name starts with a period, and therefore is regarded by Windows as a file extension rather than a file identifier. Many of these names (e.g. .options, .config) are common filenames on Unix platforms where starting a file name with a period marks it as a hidden file.

34.8% of folder names (148,536) contain one or more spaces, with 18.4% (48,494) containing a date in the filename. 14.9% (63,820) have at least one underscore, 11.1% (47,594) contain at least one hyphen, 4.8% (20,487) contain a period (in addition to the one that separates the extension), with commas occurring in only 0.8% of files (3,389).

5.3.12 Folder Names

Folder names can also be up to 255 characters in length. **Figure 154** shows the distribution of the length of folder names. There are two peaks showing on this graph. The first peak represents 4 character filenames, and the second peak represents 8 character filenames. 4 digit years are very common as 4 character filenames, as are the names 'temp', 'docs', and 'misc'. Common 8 character filenames include 'personal', 'lectures', 'articles' and 'My Music', a system created folder. Also featuring were the folders '_vti_cnf' and '_vti_pvt', which are created by Microsoft FrontPage.

The longest folder name is 220 characters long. The shortest folder names are single characters. The most common of these are the digits 0, 1 and 2.

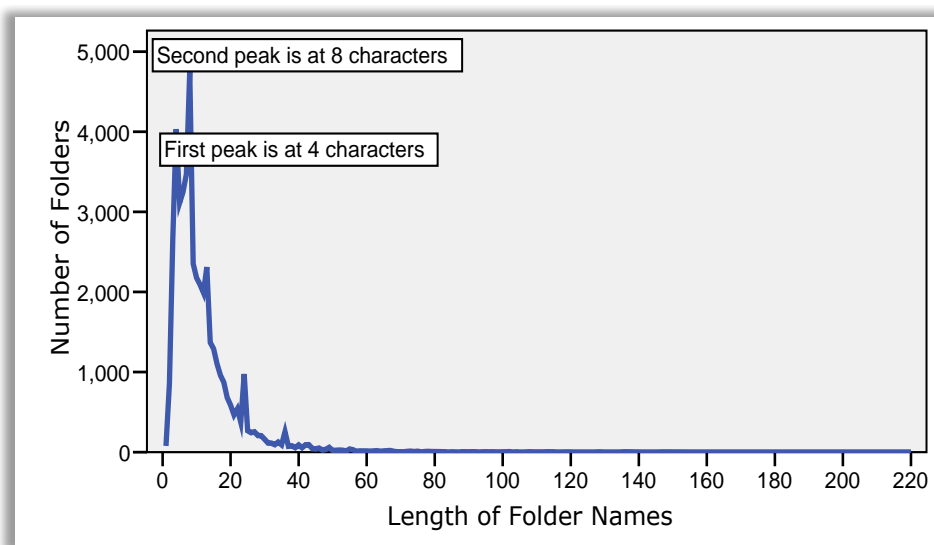


Figure 154: Line graph showing distribution of folder name length

34.6% of folder names (15,887) contain one or more spaces, 13.3% (6,117) have a date in the name. 9.7% (4,436) have at least one underscore, 6.2% (2,822) contain at least one hyphen, 2.4% (1,103) contain a period, with commas occurring in only 0.9% of files (409).

5.3.13 Versions

Many users create multiple versions of a document. As far as the Windows operating system is concerned, these are unrelated files. With such a large document corpus and no help from the file system in identifying related files, it is not possible to individually examine file names and assess whether they are part of a versioned series.

A substitute is to scan the file names and look for particular character sequences that may indicate versioning has occurred. Some people indicate versioning using the word version. This can either be preceded or followed by a description, date or numeric version indicator (e.g. "final version", "submitted version", "version 5"). In order to avoid counting file names that contain words like diversion or conversion, the string searched for was " version" (with a preceding space).

53 participants (73%) have at least one document containing the word “version” in its file name. There are a total of 2,057 files containing the word “version” (a mere 0.5% of all documents). One participant who heavily uses this technique had 1,359 versioned files. The average number of versioned files of those participants who had them was 38, with a standard deviation of 185.5.

Another common technique for versioning is to use the letter v followed by a numeric identifier: e.g. v1, v5, v23. The pattern searched for was a space, period, underscore or hyphen followed by a “v” followed by any digit. Optionally, there could also be a space, period, underscore or hyphen between the v and the digit. This eliminates matches for files with names like mov32 or cv2002.

57 participants (78%) used this technique, with a total of 2,534 files named with this pattern. The participant who uses this technique most has 359 versioned files. The average number of files versioned like this is 44, with a standard deviation of 59.7.

5.3.14 Satisfaction vs. File System Metrics

A correlation analysis was performed, investigating how satisfaction correlates with all the file system snapshot metrics. Very few items were correlated with satisfaction and when they were, the correlations were extremely weak.

Branching Factor was the most strongly positively correlated with satisfaction ($r=0.334$, $\text{sig}=0.004$), indicating that people happier with their file systems tend to have bushier systems. However, the correlation coefficient is low so this does not appear to be a strong effect. Another weak positive correlation was with the proportion of documents stored in locations other than the Desktop or My Documents folder. Thus people tend to be happier if they don’t use these locations ($r=0.326$, $\text{sig}=0.006$).

Two less significant weak negative correlations were also observed between satisfaction and the total number of files stored under My Documents ($r= -0.244$, $\text{sig}=0.041$), and between satisfaction and the proportion of files stored under My Documents ($r= -0.278$, $\text{sig}=0.019$). This would indicate that people tend to be less satisfied when they use the system provided My Documents folder for document storage.

5.3.15 Academic and General staff

One criticism of previous work on personal information management has been that studies have often been done on groups of academics. It has been suggested that perhaps academics manage their information in ways that are not typical and therefore their results cannot be generalised to wider populations (e.g. Nardi & Barreau, 1997).

This study investigated both academic staff and non-academic staff, known in the University as general staff.

Chi-square tests were performed to assess whether the Academics and General Staff systematically differ in their answers to the survey questions. There is no difference in self-reported organisation, or in satisfaction. There is also no difference in any aspect of self-reported Desktop use or use of the system created My Documents folder.

There is a difference in the way academic and general staff give names to files (Chi-square value=4.95, sig=0.026). 38% of general staff report giving file names to documents before they add content, with the remaining 62% giving file names to documents after they have created content. With academic staff, this split is 19% giving names beforehand, and 81% afterwards.

Academic staff are also more likely to report using time periods (e.g. years or semesters) in their folder names than general staff members (Chi-square value=9.16, sig=0.057). 77.5% of academics rate putting time periods in their folder names as important or very important, whereas only 52% of general staff rate this as important or very important.

Another important difference is in what people are usually searching for when they use a search function. 81.5% of academics are looking for their own files, with 4.6% usually searching for system files and 13.8% searching for files from elsewhere. Among general staff however, only 42.9% report looking for their own documents, with 16.7% looking for system files and 40.5% looking for files from elsewhere.

When they search, academics report searching on file name slightly more often than general staff (Chi-square value=9.81, sig=0.020). 78.8% of academics report always or often using file name in their search, and the remaining 21.2% report doing so sometimes. 73.8% of general staff always or often use file name, with 16.7% doing so sometimes and 9.5% reporting they seldom use the file name when they search.

General staff are more likely to use the address bar to navigate (Chi-square value=3.43, sig=0.064) with 35.7% of general staff doing so, but only 19.7% of academic staff reporting that they do this. General staff are also slightly more likely to use the back and forward buttons when they navigate (Chi-square value=3.93, sig=0.048), with 65.9% of general staff using these buttons but only 46.6% of academics reporting using them.

Academics are more likely to have multiple files for different versions of a document (Chi-square=4.15, sig=0.042). 90.3% of academics have multiple files for versions, versus 76.2% of general staff. Academics are also more likely to accidentally have multiple copies of a document (Chi-square=5.25, sig=0.022), with 55.6% of academics experiencing this compared to only 33.3% of general staff.

Academics tend to have spent longer in their jobs – an average of 10.2 years versus 4.4 years for general staff ($f=21.702$, sig=0.000). Academics also tend to have more experience working with

computers – 16.8 years compared with 9.7 years for general staff ($f=24.458$, $\text{sig}=0.000$). This is partly explained by the fact that academics tend to be older on average (Chi-square value=16.98, $\text{sig}=0.002$). The most common age group for general staff is the 20-29 age group (43.9% of general staff fall into this category). The most common age group for academic staff is the 40-49 group, with 31.5% of academics in this group. **Figure 155** below shows how the age groups vary between academic and general staff.

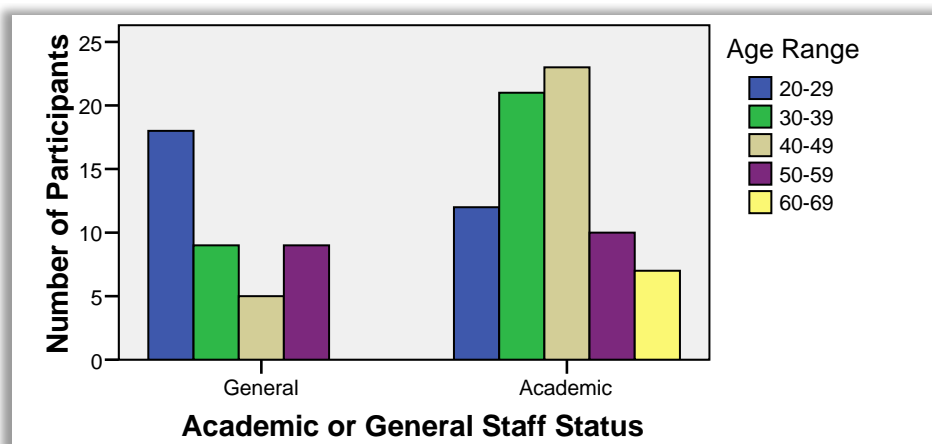


Figure 155: Bar graph showing age distribution of Academic vs. General staff

A series of ANOVA tests were performed using the file system snapshot information to see if Academic and General staff were different in the structures they create to manage their information.

On most of the file system snapshot metrics, there was no difference between Academic and General staff. There was no difference in the total number of files and folders they have, nor in the number of unique files, unique folders or empty folders. There was no difference in average or maximum depth, or in the number of shortcuts used in various places. There was also no difference in the amount of duplication (both file and folder duplication), nor a difference in measures of leafiness (number of files per folder).

There was a difference between the two groups in some measures of bushiness. There was no significant difference in the average number of subfolders per folder ($f=0.896$, $\text{sig}=.354$), however, there was a difference in the branching factor, with general staff tending to have a higher branching factor than academic staff ($f=5.41$, $\text{sig}=0.023$). General staff have a higher standard deviation of subfolders per folder ($f=4.86$, $\text{sig}=0.032$), indicating that general staff tend on average to have more uneven distributions of subfolders within their structures.

Another difference between academic and general staff was in where they tend to locate the majority of their files and folders. There is no difference in how many files and folders they tend to keep on the Desktop ($f=0.116$, $\text{sig}=0.735$). However, there is a difference in their use of the My Documents folder vs. other locations. Academics tend to keep more of their documents in the system created My

Documents folder ($f=5.19$, $\text{sig}=0.026$), whereas General staff tend to keep more of their documents in other locations (such as other folders on C drive, network drives or memory sticks) ($f=7.19$, $\text{sig}=0.009$).

5.4 DOCUMENT MANAGEMENT STRATEGIES

The interviews and snapshots revealed three strategies that the participants adopted in order to manage their document: piling, filing and structuring. The three strategies differed in the following attributes:

- Overall level of organisation (self-assessment)
- When folders are created (self-reported)
- Preferred retrieval strategy (self-reported)
- Preferred document view (self-reported)
- Use of tree (self-reported)
- Depth of structure (from snapshot)
- Breadth of structure (from snapshot)
- Unfiled documents in top level (from snapshot)
- Folders in top level (from snapshot)

All these attributes are available for the 72 survey respondents who also completed the file system snapshot. Retrieval strategy is captured in two survey questions, asking about retrieval techniques for 1 year old files, and asking about use of search tools. For the depth of the structure, average depth was used, grouped into three equal categories representing low, medium and high depth. For breadth, files per folder and subfolders per folder were both used, again grouped into three categories.

A K-means cluster analysis was performed to see if particular combinations of these attributes tended to group together, resulting in three distinct clusters. Analysis of variance indicated that several metrics were not contributing to discrimination between any clusters. These included the question on when folders are created, retrieval strategy for old files, use of tree and the breadth of the structure. These were removed one at a time and the cluster analysis repeated until all remaining variables differed significantly across the clusters. **Table 13** below shows the significance obtained for each variable.

Table 13: Analysis of Variance results for cluster analysis

Metric	F-value	p-value
How organised?	21.6	.000
Use of search	6.5	.003
Preferred view	4.9	.010
Number of Top Level Folders	67.1	.000
Number of Top Level Files	59.6	.000
Average depth	54.8	.000

Table 14 below shows the response given by the majority of participants for each cluster. Only the two most common responses in each cell are shown unless a third response was given by more than 10% of respondents in that cluster.

Table 14: Summary of cluster analysis

Cluster	1	2	3
Number of Participants	12	29	31
How organised?	67% not very 25% somewhat	76% somewhat 24% very	58% somewhat 36% very
Use of search	67% last resort 25% second choice	65% second choice 31% last resort	61% second choice 26% last resort 13% search first
Preferred view	50% list 42% details	42% list 42% details	77% details 20% list
Number of Top Level Folders	67% medium 33% high	69% high 28% medium	77% low 23% medium
Number of Top Level Files	58% high 42% medium	55% high 41% medium	77% low 23% medium
Average depth	80% low 20% medium	75% medium 25% low	54% medium 45% high

There was no significant difference between the clusters in terms of demographics. Neither age, gender nor type of work (academic or general) are related to the cluster a participant falls within.

There was a significant difference in satisfaction between the three clusters ($f=6.281$, $\text{sig}=.003$).

A discussion of these clusters and how they relate to the document management strategies identified during the interviews is deferred to the following chapter. Post-hoc tests showed that cluster 1 is significantly less satisfied with their document management than clusters 2 and 3.

5.5 CONCLUSION

This chapter has presented the results obtained from the survey. **Section 5.1** first described the design of the survey, including the development of the questionnaire from the initial conceptual model. It also described and justified the design of the contacts and procedures to be used for the survey. **Section 5.1.2** included a discussion of the steps that were taken to test the survey before delivery and to ensure the survey's validity and reliability and minimise measurement error.

Section 5.2 explored the results gained from the questionnaire. First a 'satisfaction' metric was created, expressing how satisfied people are with their document management practices. Next, the results for each section of the survey were explored to expose the range and prevalence of document management practices and attitudes. **Section 5.3** then explored the results gained from the file system snapshot, going through each metric, and where applicable, exploring the difference between reported

and actual behaviour. This resulted in the development of a classification model based on attributes posited to be important from the interviews, which showed three distinct groups of document management behaviour and structure.

The previous two chapters have presented the results obtained from both the field studies and from the survey research, and this chapter now integrates these results.

In **Section 6.1**, the results of the survey are used to validate and refine the conceptual model presented in **Chapter 4**, resulting in an updated model. **Section 6.1.2** presents a model of document management activities that any personal document management system needs to support.

In **Section 6.3**, three distinct document management strategies are presented, justified by the findings from the interviews and survey. **Section 6.4** then presents user interface design guidelines. These take the form of detailed and rich user personas based on the three document management strategies, coupled with design guidelines specifically tailored for each of these personas. In addition, general guidelines and suggestions are presented for the development of future personal document management user interfaces. Formal conclusions will not be drawn in this chapter, but will be presented in the following chapter.

6.1 CONCEPTUAL MODEL VALIDATION AND REFINEMENT

Section 4.2.2 presents the conceptual model developed from the interviews. A summary of that model was given in **Figure 22**, and is presented here in **Figure 156** for convenience.

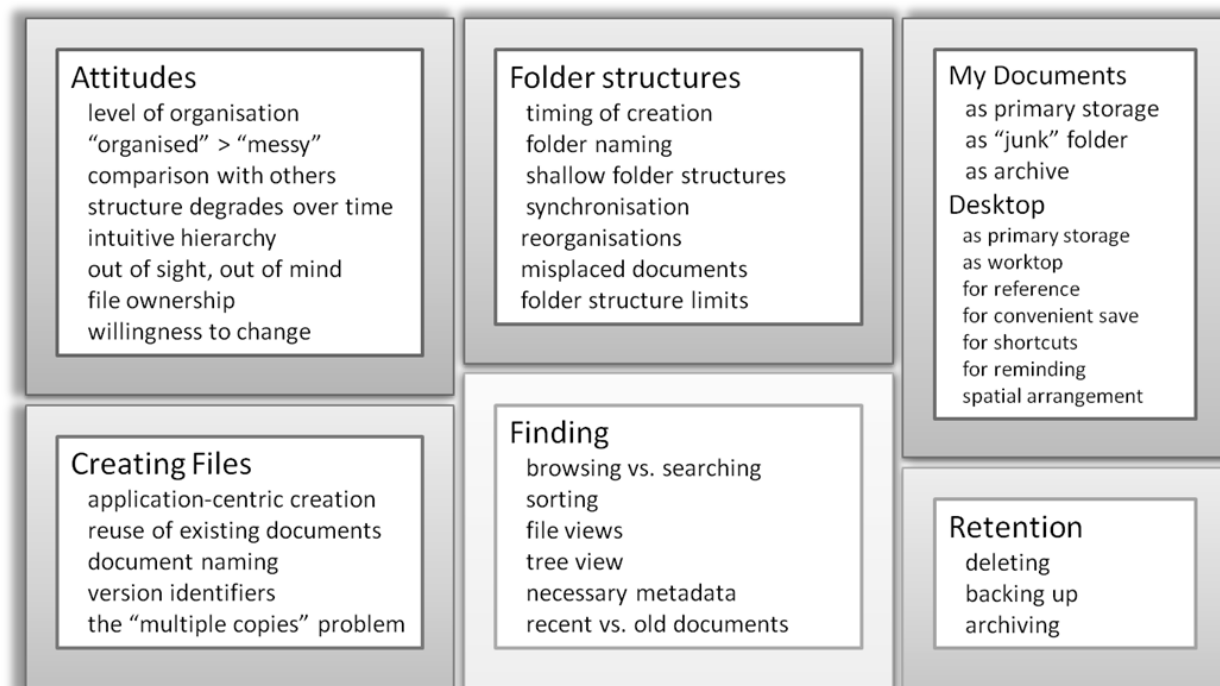


Figure 156: Initial conceptual model created from interview analysis (repeat of Figure 11)

6.1.1.1 Attitudes

Participants’ assessment of their own level of organisation is linked to their overall level of satisfaction with the personal document management. Those who perceive themselves as being more organised tend to be happier overall than those who don’t. There are two components to this, which probably reinforce each other.

First is that those who self-assess as not particularly organised do so because they know that some of their practices are not considered ‘organised’ – for instance, giving temporary file names, dumping documents in folders to be filed later, having inconsistent organising or naming schemes. These contribute to them spending more time on finding files than they otherwise would. More time spent and more frustration leads to lower overall satisfaction with their document management practices.

The second is the idea of a hypothetical ‘perfect organisation’ against which people measure themselves. Trying to attain that level of organisation was seen to be a good thing. The opposite end of the spectrum was ‘messy’, and people do not want to be considered messy. This was confirmed in the study with over 90% of people agreeing that they think it is important to have well organised documents. Since it was noted in the interviews that people often compared themselves with others, or were interested in learning how other people performed document management, it seems that people may feel less satisfied with the document management structures simply because they perceive it to be messy or poorly organised, even if the actual amount of time or effort they spend organising their documents isn’t impacted.

Least effort

One common theme running throughout the survey responses was that people want to invest the least amount of time and effort in their document management as possible. Common complaints were about the time taken to use the search tool, or to click down through the hierarchy. Feature suggestions commonly involves suggesting ways of doing something that involved less effort or less time. Other research has confirmed the general reluctance of people to spend any more time or effort than absolutely necessary (Kao et al., 2003). Users will expend the least effort they can in their document management practices.

Intuitiveness of hierarchy

The folder hierarchy is intuitive to many people and reflects the way they think about their documents. This finding from the interviews was reinforced in the survey, with the number of people in free-form comments mentioning the ability to create a folder structure and appreciating the flexibility to create their own organisation scheme within it.

Out of sight, out of mind

Some people like to have things out of their immediate view, revealing levels of details when required, but otherwise not intrusive. Others appear to like having everything displayed in front of them. Calls for an overview interface in the survey reinforce this.

File ownership

People need to feel they have control and ownership over the folders and files in the collection. This theme was reinforced in the survey with people commenting that one reason why they didn't like the system provided My Documents folder was that they didn't create it and therefore didn't have full control over it.

Willingness to change

Most people are willing to change their document management practices in order to be more organised, however some people are resistant to change. This was confirmed in the survey, with approximately three-quarters of the respondents saying they would be willing to change. Habit is a very powerful force, with several people giving habit as the reason for various document management practices they engage in. Once someone has a reliable way of doing something, they are comfortable with that and may not be willing to change unless there is a compelling reason to do so (or unless they are forced to by the change of a system).

A related theme that came up several times in the survey was participant's lack of knowledge. Many suggested the addition of features in Windows XP that were in fact already available. And several indicated that they hadn't availed themselves of available view options because they hadn't known it was possible. People don't tend to receive any training in personal document management. They are generally left to themselves to figure it out. For instance, the University of Auckland has run

professional development courses teaching people how to deal with their email, manage tasks and projects and manage their time, there are no courses teaching people how to use their documents. Basic computing courses teach the basics of creating folders, saving and opening documents, but usually don't discuss any more advanced features like changing views, sorting, advanced search options, and how to change the Desktop to enable spatial layout of items. Very few people would consider it worth spending time investigating the topic themselves, since people are generally more concerned with getting on with their tasks.

Since some people are using their document management tools sub-optimally due to lack of knowledge, one valid question is whether or not we need to change the tools or whether we merely need to train the users to use them more effectively? However principles of usability would suggest that a good software tool would not require extensive training in order to be effective – it should either be designed so it is effective without training, or it should incorporate within it training of the user as they use the system.

6.1.1.2 Creating Documents

Application-centric document creation

The interviews indicated that people mainly create documents through the appropriate application rather than through their file system. This was confirmed in the survey, with the majority of people opening the application to create a new document. Less than 10% of respondents create their files directly in Windows Explorer. Although using Windows Explorer is still important for the 27% who use it to locate an existing document to open and reuse, the majority of people name and place their files through the Save dialog boxes of applications. This is important since it means that designing a document management interface doesn't only involve creating a file management interface, but also a coherent set of Open/Save dialog boxes. This also means that any unified interface can easily be disrupted if applications are able to use their own custom dialog boxes.

Naming documents

The interviews found that firm naming conventions are rare. Although most people have some internal guidelines as to how to name particular files, these are usually only applied to certain subsets of files and are neither widespread nor consistent. The survey suggested naming conventions are more common than the interviews would indicate, with most people reporting always or often using naming schemes. In order to know the contents, people are split with half needing name or name and type, and the rest needing location also. Thus the metadata provided by the folder structure is important in resolving a file's identity for half the people. There were a few other options mostly indicating they'd need date or size or both.

The main intention when selecting a file name is that the name is something the person will be likely to remember or recognise later. Most naming conventions involve encoding metadata about the file into the filename, such as the time period it relates to, the project or course it is for. This often duplicates the names of the folders in which the file is contained, but is sometimes necessary because files often move outside their original folder context. This is not an uncommon scenario, with more than 10% of the survey respondents reporting they usually use the folder name in the doc name.

The survey confirmed that most people begin creating content in their document and give it a name later (on first save). Documents that are acquired from elsewhere rather than created will usually have their names retained if meaningful. 17% of people report that they usually rename documents received from elsewhere, and 17% usually won't.

Version identifiers

Managing multiple versions of documents in separate files is a very common practice and a source of many problems since there is no system support for this. People come up with a range of possible versioning schemes, and are not necessarily completely consistent in their use. The survey confirms that 80% of participants use multiple files for versions, and that 42% report sometimes losing track of the current version, and needless to say are less satisfied because of it. Those that report losing track indicate that they think that either the fact that they have no systematic way of versioning is responsible, or simply that they forget or are too busy to do it.

By far the most common versioning scheme is to include a version number in the file name (a practice adopted by 57% of the respondents), followed by the use of dates. Only a few people report using descriptions or moving old or current versions to another folder. A few people reported using a combination of these techniques.

The “multiple copies” problem

The interview participants indicated that having documents in multiple locations causes synchronisation problems for several participants. This was especially true with files shared between multiple computers or storage media. The survey confirmed this issue; with 47% of respondents having this problem. They indicated that some of the main reasons for it were due to multiple saving of email attachments, or forgetting they had a file and creating it again. Other reasons included deliberately making copies, for backup or archive purposes, to put a file in a location where it was easier to upload into Cecil, and because of files being transferred between computers.

6.1.1.3 Folder structures

Timing of creation - in advance, just in time, cleanup

Folders are created for a number of different reasons. They can be created before there are files to be placed within them, created ad hoc to contain files needing to be saved, or created in order to clean

up and move existing documents. Most participants reported using multiple folder creation tactics depending on the circumstances. In advance creation sometimes involves the creation of entire folder structures, often similar to or duplicating existing folder structures. The survey confirmed this, with just in time creation being the most prevalent (reported by 56% of respondents). Folder creation in response to cleanup was reported by 28% of respondents, with the remaining 16% creating in advance. Note that as suggested in the interviews, it is quite likely that many people use a combination of these techniques at different times. The survey asked which technique they would *usually* employ. The survey also found that those who create in advance tend to be happier with their file system overall.

Folder naming

Folders are named after time, topics, genre, file type, projects/tasks/courses and people. These can be combined into hierarchies in many ways. Temporary folders are often created to hold short-lived files. The survey queried how important each of these aspects was in folder naming, with the exception of genre (since few people know what document genre is). The most important aspect was judged to be using project or course, followed closely by subject or topic and then time. Purpose or use was judged to be reasonably important, while file type was judged to be unimportant.

Reorganisation

Many participants spoke of cleaning up, organising or reorganising their files. It is frequently done on a periodic basis (such as every semester or annually), but may also be done in response to rising level of mess, or continually.

Folder structure limits

Many participants remarked that reorganisation activities such as splitting a folder into multiple folders or creating subfolders were prompted by a folder reaching some limit. This is supported by the fairly consistent and low average number of files and folders people keep in their folders.

6.1.1.4 Windows XP Document locations

My Documents as primary storage

The common uses for the system-provided My Documents folder are as a primary storage location, as a “junk” folder, and as an archive for documents that are no longer active. Most of the survey participants reported keeping files in the My Documents folder. For those who don’t, the main reason is that they use another location out of habit. The use of network drives was also a factor, providing portability benefits, since the network drives are accessible from any computer on the University Network as well as through a remote desktop connection.

Desktop as primary storage

The Desktop is used for a variety of purposes. It can be used as primary storage, as an area for working files, as a location for reminding of things to do, as a location for quick access to reference

material, as a convenient dumping ground for downloaded documents, or as simply a shortcut repository. Some users do not use the Desktop at all.

They survey found that the majority of respondents don't keep documents on their Desktop (with some respondents not even knowing it was possible to do so). Of those who do use it, most don't use any folders, having just individual documents and shortcuts. Only 16% of respondents keep folders of files on the Desktop. Most participants reported that the documents on the Desktop were created by themselves. This supports the idea of it functioning as a worktop, rather than a convenient save location for downloaded documents. Slightly more than half reported that the Desktop mainly contains documents they are editing, rather than reference documents. Some participants said it was an equal mixture of reference and active working documents.

When asked why they use the Desktop, the main reason was that it is for ease of access to frequently accessed documents. A handful of participants said it was a convenient save location, and another handful said it helped to remind them of things that needed to be done. After being on the Desktop, most people tend to move documents to a permanent home or archive location. Few files on the Desktop are deleted. This suggests that Desktop documents tend to be working documents then transition to archived documents, rather than being ephemeral documents which would most likely be deleted. The length of time a document stays on the Desktop varies, with more people indicating it was a matter of months than weeks or years, and only a few people saying that the lifetime tended to be measured in days.

Some users take advantage of the spatial features of the Desktop, grouping items spatially, while others simply use the default ordering of creation time. 66% use spatial abilities.

6.1.1.5 Finding Documents

The majority of the participants reported that if they need to locate a document, they would browse to it in their folder structures. This browsing technique is also known as location based search. The survey confirmed this predominance of a tendency to browse rather than search. This cannot be construed as a clear preference in all cases, since many people weren't familiar with the ability to do full text search in Windows XP, and others complained about how slow it is.

A minority reported using search as their primary means of finding a document, with keywords from filename being the most common way of trying to locate it. Some interview participants report being unable to find a file they thought was on their system, with some duplicating work as a result. Others believe their systems work so well that this never (or seldom) happens. This was confirmed in the survey, with 62% of respondents reporting that they had completely failed to find a file they were searching for. When asked to speculate why they thought this was, the most common reason was that

it was there but they just couldn't find it. Other possible reason was that they have since removed or deleted it, or that they were wrong about saving it in the first place.

When asked about their use of a search tool, the majority of respondents said they would use a search tool if they hadn't quickly found their document through other means. More than a quarter said it would be a last resort, while those who would search first or never search were a small minority. When they searched, the majority of people were searching for their own documents, although some reported that they usually searched their systems for files that came from elsewhere. People who have to search for their own documents are less satisfied overall.

File views

Most people prefer the details view of files, a fact validated by the survey, with 49% of people using it. Some preferred the more compact list view (and some used list only because they didn't realise they could change it), and very few liked the icons view.

Sorting

Sorting proved to be a very important way of locating documents, either in search results or in folder views while browsing. Sorting by anything other than name is only possible in the details view. Changing between names, date and file type sorts were very common, with size being much less common. This fact was confirmed in the survey. For those using the details view, name was the most commonly used sort option, being used very often, followed closely by dates. File type was used as the sort criteria sometimes, with size being very seldom used. Sorting can be viewed as a sort of 'quick and dirty' way of searching or filtering within a folder. Users often change the name of their documents to force a specific sort order inside a folder.

Tree view

Many people in the interviews use the tree view when navigating, although the perceived time taken to click down through the levels is an annoyance for some. The survey confirmed this, with 70% of respondents reporting they use the tree to navigate. 65% report using the 'Up one level' button to go to the parent folder of the current folder, and 50% use the back and forward buttons to navigate between folders. Most users (77%) open each folder view in the same window, and those that do have a higher average satisfaction than those who open a separate window for each folder. Those that open separate windows tend not to be tree users.

Necessary metadata for successful finding

In order to find, most people in the interviews needed to remember something about the name of the file, keywords from the name and the type of file being searched for. Using date in a search was uncommon, but it was sometimes used to sort the results. The survey confirms these results, with filename being the most often item used in formulating search, followed by keywords from the file.

Although this had the second highest average frequency, people tended to fall into two groups – those who never use content keywords and those who did so frequently. File type is the third most common, being used sometimes on average (although there were nearly equal size groups who use it often and who seldom use it). Size was almost never used in searching for files.

Recent vs. Old documents

People who use a browse strategy are more likely to turn to search for older documents than they are recent ones. The survey showed that almost nobody reports searching for very recently used files (used within the past 2 day). While most navigated directly to the folder by browsing, one third of respondents use one of the menus of recently used files, either on the Start Menu, or on the File menu within Office applications. On average, people report they are usually able to locate their documents this way on the first try.

For older documents (one year old), 10% of respondents report that they would be able to go immediately to the exact location. 17% of people would use a search function to find it, and the remainder would browse for it. Those who report using a search function are less satisfied overall than the others who can navigate to it through their folder structure. Most people report that they are often able to find the document on the first try with their selected method, although the proportion of people who say ‘sometimes’ rather than ‘always’ is higher than for recent documents.

6.1.1.6 Retention

Deleting

While all participants have sometimes deleted files, there was a very clear division between those who were inclined to delete and those who were not. Some people prefer to remove anything they no longer need to allow more visual and cognitive space for useful items. Others prefer to keep everything just in case. The survey adds the information that most people delete because a file is temporary, or because it is an old version. A smaller number delete because they have a copy elsewhere, and some delete to save disk space. The most common reasons for deleting do not actually involve removing any important documents, but rather deleting copies of documents and ephemeral documents.

Many people do not take advantage of the ability of the Recycle Bin to recover deleted documents, either bypassing it, emptying it immediately or regularly. It may be seen as yet one more location to manage as part of their documents and therefore needs to also be kept free of ‘mess’. Those who do use it report not having to use it very frequently. This was confirmed by the survey, with only 42% of respondents using the Recycle Bin, with the remainder either emptying it straight away or bypassing it. Recycle bin users usually never have to retrieve a document, or very infrequently, usually yearly or monthly. A tiny percentage of people rely on it daily or weekly.

Archiving and Reuse

Many people had archive material, information that is no longer active, and it is important to keep this around since a sizeable number of new file creations involve reuse of old documents. 57% of survey respondents do not do anything special with material that is no longer active, just leaving it in place or moving it to a permanent archive location. The remainder either delete it outright, or more commonly, make a backup or zip file and then delete it. Two thirds of people don't even think about saving hard drive space when they make decisions about deleting files – for them it is presumably about visual and cognitive clutter rather than about preserving space.

Ability to reuse existing material from archives was very important and frequently reported in the field study. 27% of survey respondents report that their usual method of creating a file involves copying an existing document.

6.1.2 Final Conceptual Model

Figure 157 below shows the final, validated conceptual model that has been described in the previous section.

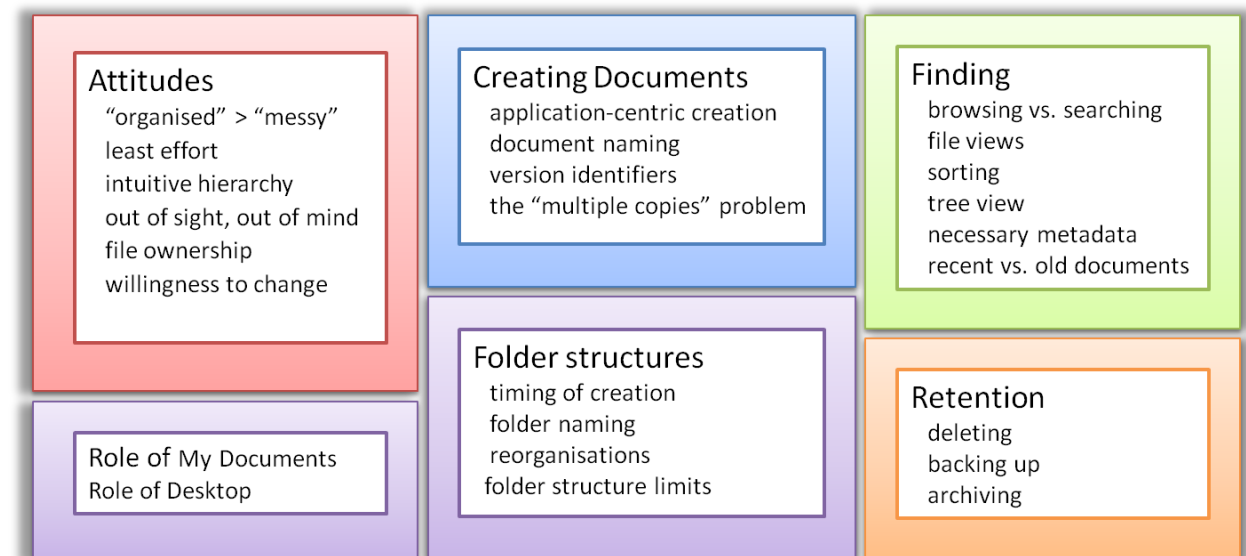


Figure 157: Final conceptual model

6.2 DOCUMENT MANAGEMENT SYSTEM CAPABILITIES

Based on the conceptual model described in the previous section, there is a set of basic operations and activities that any document management system must support. This model extends the model of PIM operations proposed by Boardman (2004) in **Section 2.1.1.1**. **Figure 158** below shows each of the operations, and the sections below describe these capabilities in detail.



Figure 158: Document management system primary capabilities

Creating/Acquiring. This is the process of adding documents to the document collection. Documents enter the collection in two ways: either they are created by the user, or they are acquired from elsewhere. Documents can be either created from scratch, or they can be a copy of an existing document. Documents can be acquired from many sources, including the web, from email attachments, and from other media such as DVDs or flash drives.

Deleting. Deleting is the act of permanently removing information from the information workspace. With modern systems having so much disk space, it is desirable that deleted items are not actually permanently removed, but instead are merely hidden from view and are recoverable.

Organising. Organising refers to the process of reorganising the information within the information workspace. This may be done in order to indicate relationships between items, or to change the display of the information. Documents can be renamed and moved to other folders, folders can be created or deleted or moved in order to contain documents, and metadata can be applied to files and folders.

Finding. Finding is an activity that involves locating one or more documents. Finding a document is never a goal in its own right – the aim is always to find a document in order to perform some other task with it. Some common reasons to find a document are to edit, to annotate and to send a document. While these activities are not supported by a document management system, the system should be aware of their occurrence. Finding does not imply only the provision of a search function, since finding can also be achieved by browsing the information space. The minimum requirement of an information workspace is only that it supports at least one method of finding information. Finding information does not change the information space.

Activity logging. Activities performed with documents such as editing, sending them to other people and printing them are not strictly speaking within the domain of personal document management. However, these document activities are part of the context of a document, and ideally should be something the document management system should be aware of.

Reminding. Documents within the collection can serve the purpose of reminding people that something needs to be done. This may be achieved purely through serendipity, when a user happens to notice something while performing another operation. Ideally however, it should be explicitly supported by the information workspace. The means used to achieve this may vary, and can include pop up alerts, displaying a document in a new location, or colour changes.

Versioning and Synchronising. Versioning means that the information workspace should be able to track updates to a document. The information workspace should be able to provide access to previous versions of a document on demand. Synchronising is the ability of the workspace to detect and integrate multiple copies of a document. While these are things the workspace should provide, they happen automatically in the background, and are not something the user interacts with every day.

6.3 DOCUMENT MANAGEMENT STRATEGIES

From the interview and survey data, it is possible to identify three distinct document management strategies. The three strategies have been named piler, filer and structurer. The piler strategy identified here is analogous to messy, no-filers, keepers, and organising neutral strategies identified by other researchers (see **Table 1: Classifications of organising strategies on page 33**). Filer and structurer are variants of the pro-organising, frequent-filer and keeper categories identified by others but have some distinct features that mean they are likely to require different user interfaces for optimal support.

Combining the results from the field studies in **Section 4.4** with the cluster analysis described in **Section 5.4** gives the following table describing the three strategies.

Table 15: Summary of document management strategy clusters

Cluster	Piling	Filing	Structuring
How organised?	Not very	Somewhat	Somewhat/very
Use of search	Last resort	Second choice	Second choice (sometimes first)
Preferred view	List/Details	List/Details	Details/List
Number of Top Level Folders	Medium	High	Low
Number of Top Level Files	High	High	Low
Average depth	Low	Medium	Medium/High

The first cluster perceives themselves as relatively disorganised, preferring a list view, with a medium number of top level folders and a high number of top level files and relatively shallow system. This suggests the file system of someone adopting the piling strategy identified in **Section 4.4.1**. The second cluster is perceived as more organised, with just in time folder creation, combination of browsing and searching only as a last resort. The structure is medium in depth and width and has a moderate number of unclassified top level folders. This suggests the adoption of a filing strategy identified in **Section 4.4.2**. Members of the third cluster have high depth, low level of unclassified files, in advance or just in time creation and consider themselves to be fairly organised. They most closely match the structuring strategy described in **Section 4.4.3**.

Some results from the classification model based on the survey data differs from the attributes that were listed for the different document management strategies in **Section 4.4**, and differs from that suggested by the existing theory presented in **Section 2.4.4**. In particular, it was anticipated users of a piling strategy would make greater use of search tools to compensate for their lack of folder structure. However, it is possible that their piling strategy means that most of the time they can browse through their top level documents, assisted by sort options until they find their target document. In this way, they are predominantly relying on a browsing technique rather than search. In contrast, adopters of a structuring strategy were not expected to be heavy users of search, since the effort they expended in structuring their folders should pay off by providing more effective browsing. However the survey results showed that structurers were more likely to search in their own documents. This result has subsequently been independently observed in a study of email (Teevan, Alvarado, Ackerman, & Karger, 2004).

It is unclear whether more frequent searches mean the document management system is less effective. It is possible that the folder hierarchy makes the search much more useful through being able to search only a related subset of the documents, and because the metadata provided by the folder path makes recognising found documents easier. However, it is clear that this research does not support the idea that there is simple trade-off between filing effort and finding effort or that filers and pilers position themselves differently on this trade-off (as was suggested in **Figure 19**). It seems that filers and particularly structurers do not necessarily benefit from their increased effort spent organising in terms of improved finding efficiency. There are clearly other benefits that structurers derive from this, such as feeling more organised and having a conceptual map of their documents. Viewed in terms of distributed cognition (see **Section 2.4.2**), it seems that structurers are using their workspace much more heavily as a cognitive aid, offloading their mental representation of their information landscape onto their filing system. Pilers may not have the need for doing this, and thus their workspace seems unstructured, but is in fact optimised for the way in which they need to interact with their documents.

Figure 159 shows how the actual trade-off curve might look compared to the curve proposed by the existing theory. More research would need to be done examining the amount of time spent in document management activities by adopters of the various strategies before a determination can be made.

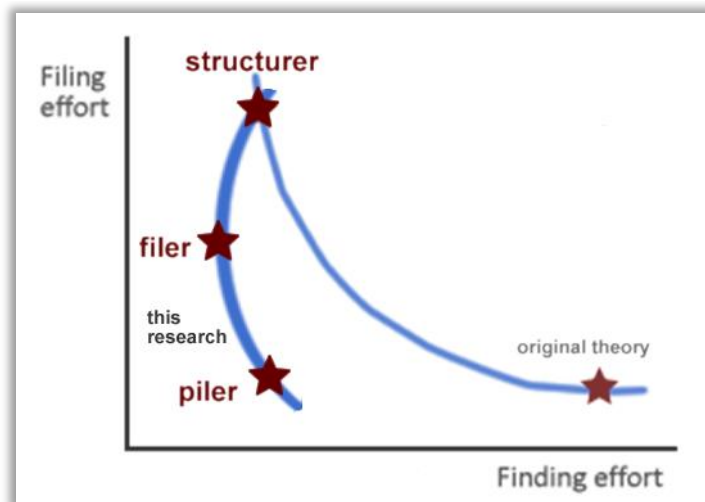


Figure 159: Possible altered trade-off between filing and finding

It was also anticipated that adopters of a piling strategy would be much less inclined to use the tree, but since the question about tree use didn't ask for frequency, there is no way of knowing whether they use it as much as users of the other strategies.

People do not necessarily neatly fit these three strategies. People will at times adopt one or the other depending on the circumstances, although there is usually having a dominant preference. These three categories collectively cover the spectrum of personal document management behaviour and therefore an interface that can accommodate all three should be useful to everyone.

6.3.1 Piling

A person adopting a piling strategy doesn't really file his documents; he just lets them pile up in various convenient locations. Folders are usually created in order to dump a large group of old documents that are no longer needed. Because folders are rarely created, the folder structure tends to be fairly shallow, with many folders and files at the top level of the structure. Because recently used files are always easily available, they are retrieved through browsing, with sorting often used to locate the most recent document. A piler may make periodic half-hearted attempts to delete things or organise them into folders, but more because he feels this is how he is supposed to do it than any perceived usefulness. It's peer pressure. Someone adopting a piling strategy tends to be a high Desktop user, since one of the key concerns is least effort and maximum availability. Minimising visual clutter

isn't really an issue, nor does he feel any need or desire to organise documents in order to get an overview of his stuff.

6.3.2 Filing

Someone adopting a filing strategy creates folders in order to split up collections of documents. They split folders up if the number of documents grows so large that they cannot easily spot items within them anymore. They tend to create folders either during cleanups or just-in-time as they need to save a folder that doesn't fit an existing category. They do have a hierarchy, although it is moderately broad and not particularly deep. They are likely to have some files in the top level (pending cleanups), and quite a few folders as well, resulting in a tree of moderate depth but high breadth. There is no particular preference for view, but they are much more likely to locate files by browsing their structures than searching. They would generally consider themselves to be relatively organised.

6.3.3 Structuring

Someone adopting a structuring strategy intensively organises their files, creating deep and meaningful document structures, often before there are documents to put in them. Related folders are grouped together into more levels of nesting, in order to hide complexity and indicate their relationship. This results in a fairly narrow and deep tree, often with fewer than 3 or 4 top level folders and very few or no files at the top level of their folder structures. They are more likely to browse through their structures although because there are so many folders to inspect, if they can't remember where something is they will readily search, particularly for older files. Browsing is often done using the tree, since the tree gives them an overview of how everything fits together. The parent folders give context to the subfolders. They get frustrated with views that don't show them the full context. For instance, search that only shows them the file name is very irritating. Showing the parent folder is even better, but they really would prefer to see the full path for context. Folders are often created in advance, as soon as a new responsibility, project, course or something appeared on their horizon, to have a place to store the documents. They tend to consider themselves very well organised.

6.3.4 Relationship between structure and strategy

It is clear from the classification model developed that users of different strategies have different structures. Since workspaces with different structures will provide different levels of support for cognitive offloading, and will have different cost structures to access information within the workspace, the structure of the workspace will necessarily influence the strategy that is adopted. Thus strategy and structure are inherently interrelated. Changes in the strategy will influence the structure and changes in the structure will influence the strategy. The two co-evolve, as shown in **Figure 160**.

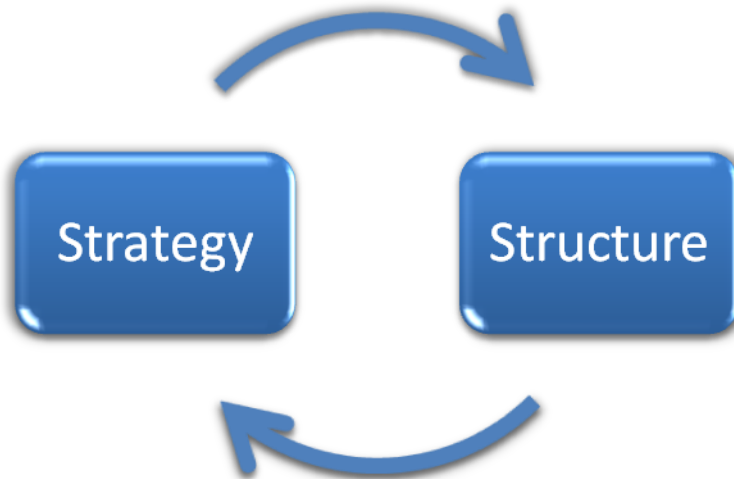


Figure 160: Relationship between strategy and structure

This is an example of the Task-Artifact cycle (see **Section 2.2**), with the artefact in this case being the file system created by the user, rather than the file system management software. The user needs and problems motivate them to create or change a certain structure to support their tasks, and the way in which they carry out those tasks will change depending on the kind of structure they have. Just as with biological evolution, it is entirely possible that there are multiple stable and successful combinations of strategy and structure, and there may not be a single optimal solution. The solution adopted may be influenced by individual differences, or on chance events such as inheriting another person's filing system. More research is necessary to identify the factors influencing the development of particular strategy/structure combinations.

6.4 USER INTERFACE GUIDELINES

This section draws on the refined and validated conceptual framework presented in **Section 6.1**, the model of personal document management software presented in **Section 6.1.2** and the three user strategies described in **Section 6.3** in order to develop guidelines for the development of user interfaces for personal document management. This section first presents three detailed user personas which together encapsulate all main user behaviours observed in this research. Following each, specific guidelines for supporting each persona are presented, followed by general guidelines that are applicable to all users of personal document management systems. A summary of these guidelines is presented in **Figure 161** below.

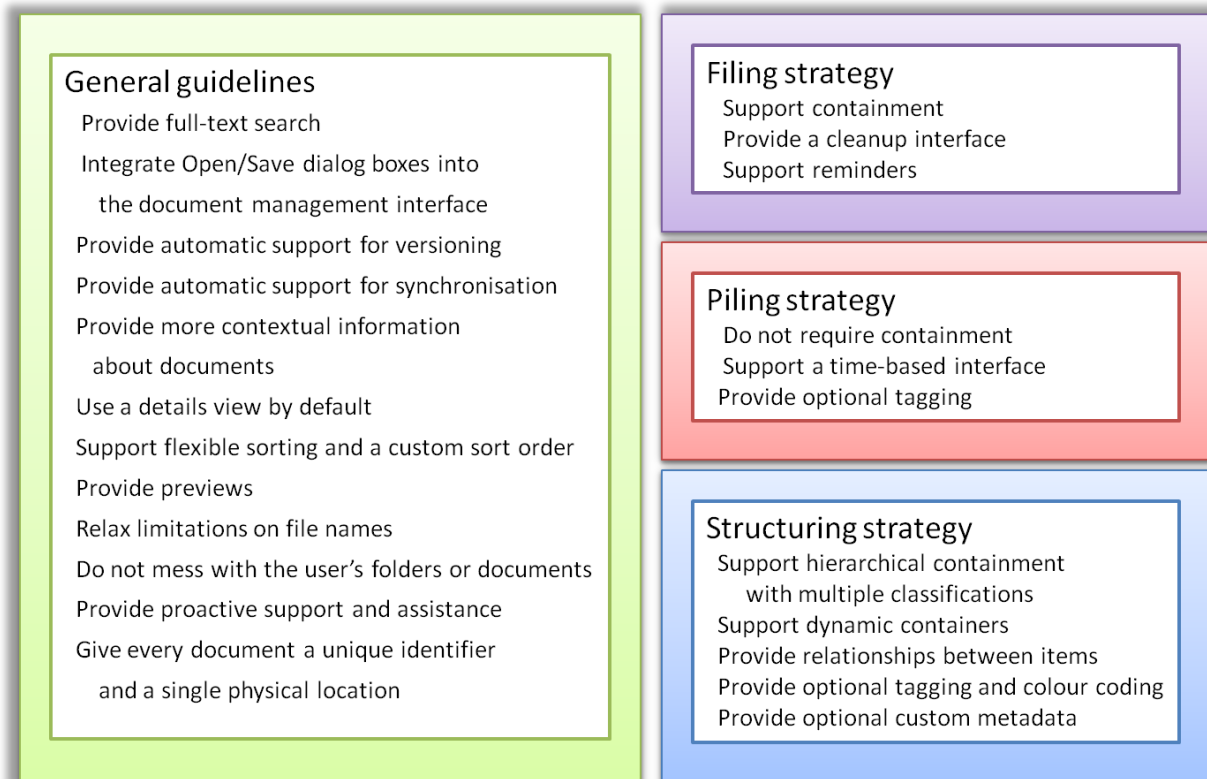


Figure 161: Summary of guidelines for development of document management user interfaces

6.4.1 Personas

Personas, as defined by Cooper are “composite archetypes based on behavioural data gathered from many actual users through ethnographic interviews” (Cooper, 1999; Cooper & Reimann, 2003). Personas provide many benefits, including providing a model of user needs, allowing differentiation between different types of users, and facilitating prioritisation of users. Personas help designers with the following tasks: (from (Cooper & Reimann, 2003 p.56):

- Determining what a product should do and how it should behave. Persona goals and tasks provide the basis for the design effort.
- Communicating with stake holders, developers and other designers. Personas provide a common language for discussing design decisions, and also help keep the design centred on users at every step in the process.
- Building consensus and commitment to the design. With a common language comes a common understanding. Personas reduce the need for elaborate diagrammatic models because, as the authors have found, it is easier to understand the many nuances of user behaviour through the narrative structures that personas employ.
- Measuring the design's effectiveness.

They also help prevent the problem of trying to design for all possible users (the elastic user) simultaneously, prevent the designer designing only for him/herself, and focus on the most important interactions, rather than edge cases (Cooper & Reimann, 2003 p.56). For personas to be useful, they need to be grounded in thorough research about the user population.

Some features of personas are:

- Personas are represented as specific individuals.
- Personas represent a class of users in context, not a particular user. The persona encapsulates a unique set of usage patterns.
- Personas have motivations and goals. These provide the fundamental information on which the design is built.

Following these guidelines, the following sections present three personas for personal document management.

6.4.1.1 Nathan (piling strategy)

Nathan works hard and plays hard and is always in a hurry. There are never enough hours in the day but he always has some Red Bull handy to keep him going. He drives a fairly old beaten up V8 Holden Commodore. His friends complain that it's always full of papers and junk (especially empty Red Bull cans), but it is certainly fast.

When he gets into the zone, he can be completely absorbed in a task for hours – all he needs is enough Red Bull and maybe some energy chocolate. It's not uncommon for him to look up from his work and discover it's 9pm already, and he's frequently late for social events because he got caught up in doing something else. He has one of the messiest desks in the office, since he just doesn't see any value in spending time and effort to file everything properly. His flat is even worse; he can barely see the floor. But as long as he can find clean clothes and anything he's looking for, he sees no real need to tidy up, especially now that he's moved out of home and doesn't have Mum nagging him to do it. After all, any time spent cleaning up is time that could be spent working or playing.

In his office he has piles of paper and books stacked everywhere, but he knows that he can always find anything he needs by going through the pile. If it's something he recently used, it'll always be near the top and it usually won't take him very long to find it. And since his life and his job moves so fast, he doesn't need to go back to old stuff very often anyway.

His computer looks a lot like his office. He usually saves everything on his Desktop because it is one of the easiest places to save things – no thought required. He likes knowing that everything is right there in front of him where he can access it quickly. After all, if he saved it, it's because he needs to do something with it and soon. When he creates files, he usually just lets the application suggest a default filename, since it requires less thought. If he's creating a document, the filename usually ends up as the document title, but sometimes when creating temporary files, he just uses the default filenames such as Document1, Document 2, Book1 and so on. He sometimes wishes he didn't have to bother giving anything names at all. One of his co-workers jokes about him being too lazy to both with proper naming and filing, but Nathan doesn't see it that way – as far as he's concerned, anything that doesn't directly affect the quality of his work isn't considered very important, and that includes filing.

He usually lets things just pile up on the Desktop until he runs out of space. When that happens, he just deals with it the quickest and easiest way he can so he can get back to work. Sometimes he deletes stuff which he is finished with and which he knows has no further use, but usually he just dumps everything except the few most recent active documents into a folder. After all, why spend time deciding which files need deleting and which should be kept? It's not as though he's running out of disk space. Sometimes he wishes that the old stuff would just disappear so he didn't even have to worry about it at all.

When he wants to find a file, he just grabs it from the Desktop. His most recent files are always on the end of the list so it'll be pretty easy to find. If he did a clean up recently, he might have to look in the latest dump folder, but usually what he needs will be on the Desktop. If something doesn't jump out at him immediately or he knows he's looking for an old file, he doesn't waste time browsing around looking for it, but jumps straight into his search tool and enters keywords from the document title. After all, if he's looking for a document, he knows what it is, and he knows that the file probably had a long descriptive title. The faster the search, the better, since all he wants is to find the document and get back to work as quickly as possible. The less thought that has to go into the process, the better.

6.4.1.2 User Interface Design Guidelines for Nathan (piling strategy)

Do not require containment

Nathan doesn't need a folder-like containment mechanism in order to group his documents, since he is interested in expending as little up-front effort as possible. This doesn't mean that folder or a grouping mechanism needs to be completely absent, just that if present, it should be optional. It should be entirely possible to use the interface without ever having to think about where to put something or in what to contain it. The 'out of sight, out of mind' principle (**Section 6.1.1.1: Out of sight, out of mind**) isn't one that Nathan subscribes to.

The attempt to take literally the piling paradigm to create a user interface that supports piles is misguided when it comes to the personal document management piler. Electronic implementations of piles (e.g. Mander et al., 1992) are a containment mechanism just like folders. Conceptually, they operate exactly as folders although with a slightly richer visual representation, one which folders views could easily match (and with picture folders starting to show thumbnails of contents, this is getting closer). The nature of the piling strategy is that followers don't really want to group and organise things. He adopts piling because it involves the least initial effort.

This doesn't mean that a containment or grouping or folder needs to be completely absent, just that if present, it should be optional. It should be entirely possible to use the interface without ever having to think about where to put something or what to contain it in.

Support a time based interface

Time based retrieval is more important to users of a piling strategy than users of other document management strategies. Although most people don't have a very highly developed sense of time (see **Section 2.4.3: Theories of Classification**), the piler naturally has (or maybe is forced to have) some sense of chronology, since their pile stacks up in order of creation/acquisition. While they don't need to remember absolute times or time spans, they need to have a relative idea how far back through the Desktop stack to look, or how many cleanup folders back to look for something (**Section 6.1.1.5 Necessary metadata for successful finding**). An interface such as Lifestreams (Freeman & Gelernter, 1996), provided it had very strong search support, would probably suit the piling strategy very well.

One way of leveraging this tendency is to ensure the default document view shows all recent files ordered by either when they were most recently used or when they were created. The Desktop could potentially use the same view, making the view easier to access. This view should be dynamic, rather than the static view currently offered by the Desktop. Items that have not been used recently should just disappear from view. Thus, the default view might show an item that was added a month ago but which was used three days ago, while an item added two weeks ago but not used since may not be visible (**Section 6.1.1.5: Recent vs. Old documents**).

Rather than having items disappear after a certain time, the view should simply show as many recent documents as possible. This takes advantage of the common practice of sorting by date to find the most recent document, and eliminates the need for periodic cleanups or dumps of files. There should be an option to 'jump back' or scroll back to show earlier sets of documents as well, giving this interface something in common with the TimeScope software (Rekimoto, 1999a), although without the spatial element.

All dates and times should be shown as relative times by default (although the option of switching to absolute times should be available), since few people have sufficiently good recall to pinpoint exactly when they created or worked with a document. Examples of relative times include '30 minutes ago', '5 hours ago,' 'yesterday' and '2 weeks ago.'

Provide optional tagging

If someone adopting a piling strategy wants to do any kind of categorisation at all in order to make sure that he is more easily able to retrieve stuff, the easiest way to support this would be to allow tags to be specified when saving the document (or added later). These can be free-form comma separated tags in which he can just type additional keywords that he might want to use to search for it but that don't appear in the document itself. This provides a way of being able to group related documents without the containment semantics, since it is easy to create a view of all documents sharing the same tag or tags. The advantage of tagging is that it lets the user add words they associate with the

documents, but which might not appear within it. This makes future searching more effective (**Section 6.1.1.5 Necessary metadata for successful finding**).

6.4.1.3 Linda (filing strategy)

Linda is a very reliable person. Her friends and colleagues know that if you ask her to do something you can safely forget about it, because she will always do anything she promises to. Every morning she gets up at the same time, makes her children's lunches and drives them to school in her old reliable Toyota Corolla. Each night, she makes sure that kids always do their homework and she's never forgotten to attend a parent-teacher meeting or to return a consent slip to her children's schools. Her house isn't super tidy – it's hard with three boys! There is always some clutter around that she hasn't gotten around to cleaning up yet, but it's always clean and has a very comfortable homey feeling. She's fairly methodical and neat, and pretty well organised - she takes a list whenever she goes shopping and she always has her Christmas shopping finished by the first week of December. At work meetings, she's always the one taking minutes, since she can be relied upon to take good notes and to remember to bring them to the following meeting. One of the secrets to her remembering everything so well is her notebook in which she writes all her tasks. She also helps herself out by trying to place things she needs to do something with in areas where she will easily see them and remember.

She tends to be a bit of a hoarder, keeping all her children's baby books and school report cards, as well as their artwork from kindergarten and school. At work she keeps books, magazines and documents from previous years, since she never knows when something might come in useful later on. She customises her workspace to suit herself, putting the documents she uses most within easy reach, and those she rarely uses on her top shelf, and in the awkward-to-access filing cabinet in the corner under the window. It doesn't bother her if things pile up a bit, but eventually every few months the size of the piles will get too high and she'll have a burst of tidying and put everything away where it belongs. Once something is filed, she doesn't look at it again unless she needs to find it for some reason. She doesn't usually reorganise or clean up material that has already been filed.

She tries to do pretty much the same thing on her computer, tending to have folders for major projects, topics or responsibilities, with all the files related to that task in the folder. Sometimes if a folder gets too big so she has to scroll a lot, she might consider splitting it. While she likes to put things in their correct folder straight away, in practice she often doesn't, saving it in a temporary location first. When she gets a few too many documents in the temporary location she'll go through and file everything properly. Sometimes at the end of a project she might get rid of early drafts or unnecessary files, but she's unlikely to revisit the folder again for cleaning purposes. Generally items enter her folders on a one-way trip.

Linda doesn't like it when the list of folders gets so long she has to scroll. It's like when her physical filing cabinet is packed full and she can't get anything more in there. She'll take some folders out and put them in the other filing cabinet in the corner, or into boxes on the high shelf. On her computer, she'll either move the folder into an archive folder, or she'll burn it onto a CD and remove it from her hard drive. That way, she doesn't have to see the folder anymore, but the information is still available in case she ever needs it or anyone ever asks her for it.

She usually sees her files and folders through her applications Open/Save file dialog views, and very rarely searches for files. Because all her file and folder names are just such common sense, she doesn't have to look very far to find things. If something was many years ago, she might have to search through her archives, but that would be relatively uncommon. Usually she can find something in a minute or two.

While she prides herself on being able to use her computer pretty well, she doesn't really like it when things change. She found it quite annoying when one of her colleagues changed her file view from the default list to details and then sorted by date while looking at something on her computer – suddenly everything had moved and wasn't in its usual place anymore. She made him change it back. She doesn't like it when things change on her computer without her explicitly taking actions to cause it, being slightly distrustful of things that happen automatically without her knowledge. But she's not a Luddite, she knows that with technology things change, and she's open to making improvements as long as she can see a clear benefit and things don't change too rapidly.

She doesn't think her folder system is anything special. She just splits things up into related groups to make it easier to find things. She's pretty sure someone else could drive her file system if she was away, because she does things that just make sense. She gets into habits of accessing things in certain ways and just uses them over and over again. She isn't really that concerned about using the latest interfaces or cool tools – she doesn't mind if what she's doing is a bit slow or not optimal. As long as it reliably works, it's fine. After all, she drives a Toyota.

6.4.1.4 User Interface Design Guidelines for Linda (filing strategy)

Support containment

Users adopting a filing strategy need a containment mechanism in order to group their files into manageable locations. The standard folder metaphor would probably work very well, although there are many other ways of implementing containment semantics, which would also work (**Section 6.1.1.5: Tree view, Section 6.1.1.1 Intuitiveness of hierarchy, Section 6.1.1.1 Out of sight, out of mind**). Different means of visualising containers should be explored. One place to start would be developing views that allow more of the hierarchy to be seen at once, since a common complaint is the time taken to click down the levels. Within containers, items should be able to be viewed with or without details,

since name is probably the most important dimension. If other dimensions are used, file type, date created and date last used would be the most useful (**Section 6.1.1.5: File views**).

While the ability to change sorting is important, there should also be a custom sort or user defined sort. In this way, filers could organise things into the exact order they wanted and know things wouldn't change. This creates a sense of stability and permanence and makes finding items through known paths easier and more reliable. It also obviates the need to change the 'common sense' file names in order to force a sort order (**6.1.1.5: Sorting**).

Provide a cleanup interface

While most of the time a user of this strategy is interacting with documents through Open/Save dialog boxes, they do want a larger view of their file structure when doing a cleanup (**Section 6.1.1.3 Timing of creation - in advance, just in time, cleanup**). During a cleanup, they are going through files in a temporary location (e.g. top level folder or Desktop) and placing them into their permanent folder home. To do this effectively they need to be able to see the list of files they are cleaning up, as much of their folder structure as possible (expanded tree view), and ideally a preview, in case they need to be reminded what the document is before they can decide where to put it. In this view it must be easy to create new folders and to reorder folder contents in the tree (**Section 6.1.1.3: Reorganisation**).

Support reminders

Being able to place documents somewhere she will be easily reminded of them would be a good feature for this strategy. A useful addition would be the ability to explicitly set a reminder on a file, which could then be used to pop up a reminder message at a certain date and time. It could also alter the appearance of files that had reminders attached so they were more visually obvious.

6.4.1.5 Matthew (structuring strategy)

Matthew likes the good things in life – good food, good friends and a good scotch whiskey. It's not about expensive or showy (he can't stand 'bling') but about quality and precision. It's one of the reasons he loves his BMW. He loves the precision German engineering, and the quality and detail that are built into every part of the car. His apartment is very minimalist, with European styling and lots and lots of cupboards. He has a place for everything and everything in its place, and all the places are hidden behind cupboards and frosted glass or concealed in staircases. All his kitchen appliances are hidden away in appliance cubbies. The only thing visible is his deluxe coffee machine both because he uses it so frequently and because he deliberately bought one with the least cluttered and smoothest exterior. Although he tries to keep everything clean and tidy all the time, in reality things can get a little bit untidy for a few days before he tidies up again.

At work he tends to have a pretty clean desk. He has several organiser boxes on his shelves and paper trays on his desk. He prefers to always have everything in its place so he knows exactly where to

find it. He has a weakness for organising systems – he just can't go past Howard's Storage World without buying something. He recently bought a deluxe labelling machine so he can put labels on all his shelves and boxes, and got a range of different coloured labels to go with it so he can use colour codes. His friends sometimes tease him about being excessively anal about filing.

Matthew likes to be just as organised on his computer. His Desktop is clean – just shortcuts to the one or two items that he accesses all the time, which he changes depending on what he is currently working on. He tries to give every file a descriptive name, sometimes with codes to indicate the year, project or task, and placing it in its proper place in the hierarchy.

Whenever he starts a new project, one of the first things he does is select a name for it and create a folder structure (often similar to previous projects). If he knows in advance what documents he'll need to create, he'll often create the outlines for those in advance, making sure they all use the same template so all the styles and formatting are consistent. If he knows he'll be corresponding with others about it, Matthew will make a folder in his email system for it, and if he knows he'll be doing a lot of web searches, he'll make a folder in his browser's favourites to store all the related web links. If he's going to be working with paper documents or books, he'll create a label on his shelves to contain the related material.

Matthew sometimes switches between different ways of organising things – it's important to be organising things as well as he can and he's never sure if he's doing things the best way. He wonders, should he keep his trip expense reports in a separate expense report folder or should he put them in the folder with the rest of the information about the trip? He wishes he could have things in more than one place. And with some of his projects now spanning multiple years, he's never sure whether he should create year folders inside project folders or the other way around. He currently has year folders as the top level, because he stumbled across an article on the web which argued that this was more efficient.

He likes the fact that the tree view gives him an overview of the structure of his project but he wishes it was more useful, like letting him know which parts he still needed to work on and which were completed. He downloaded a trial shareware application that let him colour code his folders, which he liked for a while, but didn't like quite enough to pay \$50 for it. He's also downloaded a couple of shareware applications that present different views of his folder structures, showing how the parts of the structure are related to each other and letting him follow links from one folder to another.

When someone else sees Matthew's folder structures he feels that they aren't just seeing the places where he stores his files, they're seeing the structure of his mind. He just wishes he could make his folders a little bit more expressive of his own mental representations.

6.4.1.6 User Interface Design Guidelines for Matthew (structuring strategy)

Followers of a structuring strategy need the ability to express containment just as filers do, but they also need richer containment semantics.

Support hierarchical containment with multiple classifications

Systems must provide the ability to create hierarchies of containment, since many people appreciate the ability to create folder structures (**Section 6.1.1.1: Intuitiveness of hierarchy**). Multiple classifications enable a document to live in more than one location. Previous means of approximating this such as shortcuts or copies are not sufficient – the document actually needs to have one location but appear in multiple locations. Regardless of the location from which the file is viewed and accessed, any changes to the document or its metadata should be immediately effective in all locations. When a file is deleted, if it exists in multiple locations the user will need to be prompted whether the file should be deleted from that location only or from all locations.

Allowing multiple classifications overcomes the cognitive difficulties caused by trying to find a single un-ambiguous category for a piece of information, and allows users to include more contextual information in their categorisations (see **Section 2.4.3: Theories of Classification**). This is particularly important for personal document management, as due to the personal nature, it is very context-dominated.

A user interface should support collapsing or hiding of levels of information, to enable the ability to see an overview and drill down to detail on demand. (**Section 6.1.1.1: Intuitiveness of hierarchy**).

Support dynamic containers

Providing dynamic containers is another way of providing some of the same functionality as multiple classifications. Dynamic containers don't have a predefined set of contents, but rather display the contents based on a search. The containers in the Presto system (Dourish et al., 1999a) are an example of this, as are Outlook 2003's Search Folders. For instance, Matthew's expense reports could be stored in the folder with the rest of his trip information, but he could create a dynamic folder that presents a view of all his expense reports together. The dynamic folder can be organised into folders like any other folder.

Provide relationships between items

To a structurer, the file system is more than simply a place to store things; it is a representation of the structure of his information (**Section 6.1.1.1: Intuitiveness of hierarchy**). For this reason, the ability to make arbitrary relationships between things would be a useful extension. This can be partly automatic and partly manual. For instance, the system could track which documents are opened with other documents or emailed together with other documents and therefore infer relationships between documents. This could be presented by having a 'Related items' panel that displayed the other

documents related to the currently selected document, enabling them to be quickly accessed. In addition, there should be an ability to manually create relationships between items, thereby choosing the items that appear in the 'related items' view.

Provide optional tagging and colour coding

Other methods to provide the structuring filer with richer abilities to organise files include allowing the ability to tag documents or files with keywords (as described for Nathan in **Section 6.4.1.2**), and to colour code files and folders. These should be entirely optional but if used are entirely user-generated. The structurer can use any colours they want, and can assign an optional descriptive label to the colour, or just simply use the colour.

Provide optional custom metadata

The 'Rolls Royce' of systems for a structurer would be to allow them completely free rein to construct their own properties to be added to files and folders and to use these properties to create dynamic folders and hierarchies. These could then provide the basis of a customised search function that provides a means of finding information by arbitrary metadata (**Section 6.1.1.5: Necessary metadata for successful finding**). Whilst this provides the ultimate in flexibility, it requires considerable effort and overhead to maintain, and it must be acknowledged that relatively few users are interested in organising to quite this extent.

6.4.2 General User Interface Guidelines

Provide easy to use, fast, powerful full text search

All users rely on search tools to sometimes locate documents and thus need a very fast and robust full text search (**Section 6.1.1.5**). Although users of piling and filing strategies don't rely heavily on search tools for accessing their documents, they do use it sometimes, particularly to find old documents or documents in their archives. It needs to be as easy to use as possible, with no complicated criteria or Boolean logic. Ideally it should be as easy as a Google search.

The default sort order of results should be some combination of relevance and time. Details about the file should be shown by default, including the creation and modified dates, name and file type. It should be easy and fast to refine the search, including changing the sorting to sort on name or file type. It should also be easy to filter based on either date ranges or file type (**Section 6.1.1.5: Necessary metadata for successful finding**).

Integrate Open/Save dialog boxes into the document management interface

All users interact with their document collection through applications' Open/Save dialog boxes, and filers perform most creation, acquisition and locating activities this way (**Section 6.1.1.2: Application-centric document creation**). Thus, these need to be considered first class citizens in a personal

document management user interface. They should present exactly the same interface the user would normally use to access their files, including any preferences for views, sorting or other customisations.

For filers and structurers this dialog will probably need to be much larger than they currently are in order to provide a useful view of the file system. For adopters of a piling strategy, the dialog should be as minimal as possible with perhaps simply a field to specify the filename (which ideally should default to something sensible suggested by the document). Since pilers don't usually specify a place, there is no need for a large view of a folder structure for them to select one. There should be an option to switch views, since most users do not operate exclusively according to type.

Provide automatic support for versioning

Users should not have to create multiple files for multiple versions of a document. The file system should handle that automatically. Every time a file is saved, it should automatically create a new version with the current information. The document management interface should always show the most recent version of a file, but the previous versions should be accessible on demand. It is not necessary that they be easily or quickly accessible, since going back to previous versions is not a common activity, but it should be possible to do so when necessary (**Section 6.1.1.2: Version identifiers**). This should eliminate one major source of version problems. The other is dealt with by the synchronisation support described in the next section.

Another aspect of this is that it should be possible to mark a document or folder as being locked or archived, which essentially makes the document Read Only. Views of the document should visually represent this in some way. This provides some support for task management, since people can essentially mark their document as being complete, as well as providing a visual distinction between working and archived information. (**Section 6.1.1.6: Archiving**)

Provide automatic support for synchronisation

When a file is copied onto the hard drive, it should be checked against the files that already exist to check whether it is a duplicate (**Section 6.1.1.2: The "multiple copies" problem**). Windows XP currently does this if only if the file is copied into the same folder. It should also happen if the file is copied into any other folder. The user should be prompted about how to handle this. The options are:

- Keep the document in both locations in the file system. While there is only one document, it is accessible from both places. If any of the document metadata has changed it will be merged, with the user prompted to resolve any conflicts.
- Keep the document in one location in the file system. In this case, it will be removed from the other folder, and will keep only the metadata of the retained location.

This method should also be able to resolve conflicts between versions that occur because updates to a document were made on another system and then brought back to this one. The system can identify

when a file is saved that it is a later version of an existing file and merge the new version into the file. Since all previous versions are retained, there is no loss of information that occurs in this process.

Provide more contextual information about documents

A great deal of information about documents and what happens to them isn't being collected, and could provide rich information to assist with searching and decision making in certain circumstances. For instance, the system should record every time a document is printed, emailed as an attachment, converted into another form (e.g. Word document saved as PDF) or used as the basis of another document. This information provides more context about the document that can be useful when searching for (for instance) 'the document I sent to John last week' or 'the report I printed for the boss last month' (**Section 6.1.1.5: Necessary metadata for successful finding**).

Use a details view by default

Details view provides the most information to a user in identifying and working with their documents. As this is used by the majority of users, it should be the default (**Section 6.1.1.5: File views**). It should however be possible for users to choose to see less information if they wish, or to see more or customise the view if they want. However, given that most people do not undertake extensive management of their document management user interfaces, the most useful option should be selected as the default. The details view should include name, file type (either a text description or an icon to distinguish) and the date the file was last used, and the date it was created. Relative dates should be preferred over absolute timestamps.

Support flexible sorting and a custom sort order

Sorting is a very important mechanism used to locate documents, and sorting on any visible attribute should be easy to accomplish. In addition, it should be possible to specify a custom sort for a folder or container, in which the user can reorder folders and documents to appear as they wish. This should be remembered so that if the user switches to another sort order, they can switch back and have their custom sort presented again. This should prevent people from using file naming techniques to force documents to sort in particular ways (**Section 6.1.1.5: Sorting**).

Provide previews

To assist users in locating the document they are looking for (whether they are searching or browsing) there should be a document preview available which shows at least the full first page of a document in a large enough size for text to be read. Ideally this would also allow scrolling through the document, and if the user only wanted to check the contents rather than edit it, that could be accomplished through the preview alone without having to open the document's application.

Relax limitations on file names

File names should be unlimited length and should be able to contain any characters including slashes, question marks quote marks, asterisks and all the other characters that are currently prohibited. The

user should be free to name their file however they choose, including not specifying a file name at all. (**Section 6.1.1.2: Naming documents**)

Do not mess with the user's folders or documents

Users need a sense of ownership over their files, and so the general principle is that the system should not interfere with their structures unless absolutely necessary (**Section 6.1.1.1: File ownership**). For instance, the system should not create the pseudo-folders My Music and My Pictures. Rather, the user should be allowed to create as many folders for their pictures and music as they want, wherever they want and name them however they want. They should be able to select a custom view (pictures view or music view) for those folders, and this custom view could also be reused in displaying search results for the appropriate type of file. Likewise, the system should not move documents around or take any actions without the user's knowledge and consent⁵. User settings related to the operating system and applications should be stored elsewhere, either a designated settings folder for each user or in the Registry. These should not be intermixed with user folders and documents.

Provide proactive support and assistance

The system should attempt to provide sensible defaults for everything it can, relieving the user from having to make as many decisions as possible. For instance, the system can suggest default filenames and default tags, although these should be easy to override by the user if not wanted. The system can also have a 'teaching' interface that can (infrequently) let people know about improvements they can make to their practices (**Section 6.1.1.1: Willingness to change**). Some work needs to be done in order to work out the best time to provide this support. If the system could detect a reorganisation in progress, this would be a great time, since during other tasks such as file saving/editing, people are likely to be more focussed on the task than on their document management. A system such as this could allow people to make incremental improvements in their document management over time, enabling them to more effectively use the system, as well as increasing their feelings of control and satisfaction.

Give every document a unique identifier and a single physical location

This last guideline is not so much a user interface recommendation but a suggestion for developers of document management user interfaces. Every document should have a globally unique identifier. This will enable automatic synchronisation and version resolution to be performed much more easily, and will make it much easier to manage relationships between documents. It will also allow the metadata to be more easily separated from the document.

⁵ Note that the pile's recent folder view in which older messages 'disappear' does not actually contravene this principle. The files are not actually moved anywhere, they just no longer appear in the main screen, in a similar way to how old messages in a full inbox may not be seen as they scroll off the screen.

Each document should exist in only one location on disk, although views of that document may appear in multiple places in the document management system. This prevents versioning and synchronisation issues that can occur when a document exists in multiple copies.

6.5 SUMMARY

This chapter has integrated the findings from the interviews and the survey to form a theory of personal document management. **Section 6.1** presents a conceptual model of document management, setting out the major concepts and concerns discovered in this research. **Section 6.1.2** presents a model of document management activities that any personal document management system needs to support, and **Section 6.3** describes three distinct document management strategies. **Section 6.4** then draws all these elements together to presents user interface design guidelines. These take the form of detailed and rich user personas based on the three document management strategies, coupled with design guidelines specifically tailored for each of these personas. In addition, general guidelines and suggestions are presented for the development of future personal document management user interfaces. The following table summarises these guidelines and identifies the section or sections of the conceptual model that provides their justification.

Table 16: Summary of user interface guidelines and their justifications

Guideline	Justification from conceptual model
Nathan (piling strategy)	
Do not require containment	Attitudes > Out of sight, out of mind
Support a time based interface	Finding > Necessary metadata for successful finding Finding > Recent vs. Old documents
Provide optional tagging	Finding > Necessary metadata for successful finding
Linda (filing strategy)	
Support containment	Attitudes > Intuitiveness of hierarchy Attitudes > Out of sight, out of mind Finding > Tree View Finding > File Views Finding > Sorting
Provide a cleanup interface	Folder structures > Timing > Cleanup Folder structures > Reorganisation
Support reminders	Role of Desktop
Matthew (structuring strategy)	
Support containment with multiple classification/dynamic containers	Attitudes > Intuitiveness of hierarchy Attitudes > Out of sight, out of mind Finding > Tree view
Provide optional relationships between items	Attitudes > Intuitiveness of hierarchy Finding > Necessary metadata for successful finding
Provide optional tagging and colour coding	Finding > Necessary metadata for successful finding Finding > File Views

Provide optional custom metadata	Finding > Necessary metadata for successful finding Finding > File Views
General	
Easy to use, fast powerful full text search	Finding > Necessary metadata for successful finding
Integrated Open/Save Dialogs	Creating Documents > Application centric document creation
Provide automatic versioning support	Creating Documents > Version identifiers Retention > Archiving
Provide automatic synchronisation support	Creating Documents > “Multiple copies” problem
Provide more contextual information	Finding > Necessary metadata for successful finding
Use details view by default	Finding > File Views
Support flexible support & custom sort	Finding > Sorting
Provide previews	Finding
Relax limitations on filenames	Creating Documents > Naming Documents
Provide autonomy over use folders	Attitudes > File ownership Role of Desktop Role of My Documents
Provide proactive support & assistance	Attitudes > Willingness to change
Give every document a unique ID and single location	[Necessary to implement automatic versioning and synchronising support]

Chapter 7 presents the conclusions and describes the contributions made to knowledge as a result of this research.

7.1 SUMMARY OF RESEARCH PROBLEM AND APPROACH

The aim of this research has been to improve the knowledge about personal document management in order to inform the design of future document management and personal information management tools. Since knowledge workers today spend so much time working with information, having improved interfaces for these activities is important for ensuring higher productivity and an enhanced working environment.

This research focussed on the problem of individual management of digital documents in a work setting. While, there is a growing body of research on the wider field of personal information management, previous research on personal document management was limited in scope, and was conducted using interfaces a generation behind those currently in use today. Much of the research on user interfaces for document management was speculative and exploratory in nature – problems were hypothesized based on anecdotal evidence, and new interfaces were proposed for document management without a thorough understanding of the user requirements.

The aims of this research were:

To improve understanding of document management. In particular, empirical data was required about the document structures people create, the processes people use to manage those document structures and the problems people encounter in document management. The intention was to

develop a model of document management concepts and practices that could then be used to inform user interface development.

To develop guidelines for the design of user interfaces to support document management.

Another objective was to develop guidelines for improved document management workspaces. These guidelines are directed at designers of user interfaces and provide specific principles to follow when designing new interfaces.

In order to achieve these aims, a two stage research design was adopted, consisting of interviews and a file system snapshot followed by a survey. In both stages, both objective and subject data were collected – objective data about document structures, and subjective data about reported user behaviour and attitudes.

The **interviews and file system snapshots**, reported in **Chapter 4**, investigated personal document management behaviour with 10 participants. This exposed a range of attitudes and behaviours and enabled a rich understanding of people's feelings about and motivations in dealing with their documents. This study allowed identification of the key areas that were then explored in less detail but larger numbers in the next phase. It also included the initial development of a data collection tool to take a comprehensive snapshot of a user's file system.

The **survey**, reported in **Chapter 5**, sought to obtain data from a larger number of participants, including academics and a range of professional, clerical and managerial positions. This complemented the interviews by providing more insight into the spread of various practices and attitudes across a knowledge worker population.

The **analysis**, reported in **Chapter 6**, synthesized the richness of the qualitative interviews and the quantitative data and presented a conceptual model of document management, proposed capabilities of document management systems, a set of document management strategies and detailed user interface development guidelines including personas.

7.2 CONTRIBUTIONS

The following table summarises the contributions to knowledge made by this research.

Table 17: Contributions to knowledge made by this research

Contribution	Type	Chapter
Personal document management definitions and concepts	Theoretical	Chapter 2
Review of personal document management literature and theory	Theoretical	Chapter 2
Development of file system snapshot software, file system metrics and snapshot analyser software	Methodological	Chapter 4
Development of a personal document management questionnaire, including a metric for document management satisfaction	Methodological	Chapter 5
Collection and analysis of empirical information about personal document structures	Empirical	Chapter 4 & Chapter 5
Development of a validated conceptual model of document management	Theoretical	Chapter 4 & Chapter 6
Identification of primary capabilities of document management systems	Theoretical	Chapter 6
Identification and discussion of three document management strategies	Theoretical	Chapter 6
Provision of guidelines for developers of document management user interfaces	Theoretical Practical	Chapter 6

The following sections describe these contributions in more detail.

7.2.1 Contributions to understanding personal document management

The first contribution made to improving the understanding of personal document management was the critical review of literature in **Chapter 2**. This review was performed in three parts. The first part examined empirical studies of personal information management, with a particular emphasis on personal document management, and noted a lack of recent studies in this area. The second part critically reviewed commercial systems and prototypes for supporting personal document management. The Windows XP file management tool, Windows Explorer was described in detail, as this is the most widely used document management system in current use. Research prototypes were found to generally lack grounding in actual document management practices, and suffered from a lack of systematic evaluation. The third part synthesized theory related to personal document management, noting a growing consensus about broad information management strategies which can be conceptualised as a trade-off between effort expended in filing and effort expended in finding. The effort depends on the cost structure of the information workspace, which consists of external representations of internal knowledge structures (according to the theory of distributed cognition). In

document management, these external representations frequently take the form of a taxonomy, and so theory relating to classification was discussed.

Another contribution is methodological, through the development of a new technique for investigating personal document management. This combination of a survey and a snapshot of user's file system enabled subjective information about document management strategies to be combined with objective information about document structures. The questionnaire was developed specifically for this research and can be reused or modified for future studies, and included a set of questions which can be used to produce a measure expressing how satisfied a person is with their overall document management practices. In addition, the file system snapshot software was developed and used, along with the tools to extract and analyse the data. Because this was a new approach, a variety of metrics were developed to describe user's file systems. These can be used in future studies for comparative purposes.

The primary contribution is empirical, through the collection of information about document management practices and structures on a larger scale than has been attempted in the past. This provides quantitative information about the sizes of the document collection, the nature of the hierarchical structure, and the use of various system features such as the Desktop, shortcuts, search tools and viewing and sorting options. This information can be used (and has been used) by researchers designing further studies in this area, or by developers of personal document management systems seeking to understand the nature of the collection they are supporting.

A theoretical contribution is the development of a validated conceptual model of document management. This model sets out the major theoretical findings related to how people create files and folders, structure them, make decisions about retaining them and using various features of the document management system, and major attitudes to personal document management. This conceptual model is thoroughly grounded by a detailed qualitative analysis of people's explanations of their document management practices. It is validated by the larger scale subjective and objective data obtained in the survey.

7.2.2 Contributions to user interface guidelines

A practical contribution is a model showing the primary capabilities that any personal document management system must have. This model is based on user's explanations of what they do with their document system and their suggestions for what they would like it to do for them.

A contribution that is both theoretical and practical is the identification of three document management strategies. These three strategies of piling, filing and structuring were derived qualitatively from the interviews and quantitatively confirmed using the survey data. As these three

strategies approach the task of document management differently, the structures they tend to create also differ, resulting in distinct tool support requirements.

The ultimate contribution is the provision of guidelines for developers of document management user interfaces. These take two forms. The first is a set of three rich user personas which can be used by designers as a tool for developing requirements and assessing the usability of document management systems. The second is a set of detailed guidelines, including general guidelines that will improve document management user interfaces for all users and guidelines giving requirements for the specific document management strategies.

7.3 LIMITATIONS AND FUTURE WORK

In this study, the author acted as the sole researcher, performing all studies, tool development, data collection and analysis, and consequently there were a number of limits placed on the research scope for pragmatic reasons.

The organisation selected for the study was a research university. Although managers, professional and clerical staff were included along with academics, research in other settings would be useful to confirm the generality of these results, and perhaps to investigate the degree to which special requirements exist for particular domains.

The unit of analysis selected was an individual and their primary work computer. The interviews surfaced instances of people working in multiple locations and wrestling with the problem of replicating data from one location to another. A comparison of the structures people keep at home and at work would be very useful to see what differences and similarities there might be, as well as to explore any problems this replication can create.

Additionally, while this research identified three strategies adopted with respect to document management, additional work could examine other potential influences on these strategies. Since demographics seemed to play little role in determining the strategies employed, additional research that used the same questionnaire and snapshot but additionally collected information about personality type (such as Myers-Briggs type indicator) and cognitive ability would be very useful in understanding how these strategies come about.

This study was limited to users running Microsoft Windows XP. Although this is the dominant operating system, it would be useful to study people using other operating systems, such as Apple Macintosh and variants of Linux to see whether there is any systematic difference. Also, a useful follow-up would be to examine people using the newer Windows operating system, Windows Vista. This operating system offers much improved search capability, and a longitudinal study investigating how this affects people's document management practices would be very useful.

This research considered how people organise their documents and focussed on the strategies they adopt and the structures they create using a cross-sectional study. A longitudinal study of document management practices would be a useful complement to this and would provide additional information on time spent in document management activities as well as examining how structures and strategies co-evolve with time. It would also be able to provide information about patterns of document use over time, and about document lifetimes and access patterns.

There is scope to perform further analysis with the data collected in this study, particularly to exploit the possibilities of creating richer models of user behaviour. In particular, a fruitful avenue for further work would be the creation of models to predict satisfaction and performance for each of the three personas identified here.

This work was empirical and theoretical in nature but the aim of this research is to inform design of user interfaces. The most useful future work to continue this body of work would be to take the personas and design guidelines developed in this research and use them to develop and evaluate new interfaces for personal document management. Hopefully this research will contribute to the creation of new tools and systems that increase the productivity of knowledge workers as well as making their lives just a little bit easier.

- Abrams, D., Baecker, R., & Chignell, M. (1998, April 18-23, 1998). *Information Archiving with Bookmarks: Personal Web Space Construction and Organization*. Paper presented at the CHI'98 Conference on Human Factors in Computing Systems, Los Angeles, California, USA.
- Alavi, M., & Carlson, P. (1992). A Review of MIS Research and Disciplinary Development. *Journal of Management Information Systems*, 8(4), 45-63.
- Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, 25(1), 107-136.
- Apté, C., Damerau, F., & Weis, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3), 233-251.
- Aronson, J. (1994). A Pragmatic View of Thematic Analysis. *The Qualitative Report*, 2(1).
- Baecker, R., Booth, K., Jovicic, S., McGrenere, J., & Moore, G. (2000, November 16-17, 2000). *Reducing the Gap Between What Users Know and What They Need to Know*. Paper presented at the CUU'2000 Conference on Universal Usability, Arlington, Virginia, United States.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.
- Bälter, O. (1997). *Strategies for Organising Email*. Paper presented at the Proceedings of HCI on People and Computers XII.
- Bälter, O. (2000, April 1-6, 2000). *Keystroke Level Analysis of Email Message Organization*. Paper presented at the CHI'2000 Conference on Human Factors in Computer Systems, The Hague, Amsterdam.
- Barreau, D. K. (1995). Context as a Factor in Personal Information Management Systems. *Journal of the American Society for Information Science*, 46(5), 327-339.
- Barreau, D. K., & Nardi, B. A. (1995). Finding and Reminding: File Organization from the Desktop. *SIGCHI Bulletin*, 27(3), 39-43.
- Bellotti, V., Ducheneaut, N., Howard, M., Neuwirth, C., & Smith, I. (2002, June 25-28, 2002). *Innovation in Extremis: Evolving an Application for The Critical Work of Email and Information Management*. Paper presented at the DIS 2002 Conference on Designing Interactive Systems, London; England.
- Bellotti, V., & Smith, I. (2000). *Informing the Design of an Information Management System with Iterative Fieldwork*. Paper presented at the DIS'2000 Symposium on Designing Interactive Systems, New York, USA.
- Bergman, O., Beyth-Marom, R., & Nachmias, R. (2003). The User-Subjective Approach to Personal Information Management Systems. *Journal of the American Society for Information Science and Technology*, 54(9), 872-878.
- Boardman, R. (2004). *Improving Tool Support for Personal Information Management*. Unpublished Doctoral Dissertation, Imperial College, London, England.
- Boardman, R., & Sasse, M. A. (2004, April 5-8, 2004). *"Stuff Goes into the Computer and Doesn't Come Out" A Cross-tool Study of Personal Information Management*. Paper presented at the CHI'2004 Conference on Human Factors in Computing Systems, Vienna, Austria.
- Boardman, R., Sasse, M. A., & Spence, R. (2003, June 22-27, 2003). *Too Many Hierarchies? The Daily Struggle for Control of the Workspace*. Paper presented at the HCI International'03 International Conference on Human-Computer Interaction, Crete, Greece.
- Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in Information Systems Research: A state-of-the-art assessment. *MIS Quarterly*, 25(1), 1-16.
- Bourque, L. B., & Fielder, E. P. (1995). *How to Conduct Self-Administered and Mail Surveys*. Thousand Oaks: Sage Publications.

REFERENCES

- Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: A guide to classification and its consequences*. Cambridge, Massachusetts, USA: MIT Press.
- Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, California, USA: Sage Publications.
- Buckland, M. K. (1997). What is a "Document"? *Journal of the American Society for Information Science*, 48(9), 804-809.
- Bush, V. (1945). As we may think. *Interactions*, 3(2), 35-46.
- Carroll, J. M., Kellogg, W. A., & Rosson, M. B. (1991). The Task-Artifact Cycle. In J. M. Carroll (Ed.), *Designing Interaction: Psychology at the Human-Computer Interface*. Cambridge: Cambridge University Press.
- Cockburn, A., & McKenzie, B. (2001, March 31-April 4, 2001). *3D or not 3D? Evaluating the Effect of the Third Dimension in a Document Management System*. Paper presented at the CHI'2001 Conference on Human Factors in Computing Systems, Seattle, Washington, USA.
- Cole, I. (1982). *Human aspects of office filing: Implications for the electronic office*. Paper presented at the Human Factors Society Annual Meeting, Seattle, Washington, USA.
- Cooper, A. (1999). *The Inmates are Running the Asylum*. Indianapolis, Indiana, USA: Sams.
- Cooper, A., & Reimann, R. (2003). *About Face 2.0*. Indianapolis, USA: Wiley Publishing.
- Corbató, F. J., & Vyssotsky, V. A. (1965). *Introduction and Overview of the Multics System*. Paper presented at the Fall Joint Computer Conference.
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Crawford, S. D. (2004). What they see is what we get: response options for web surveys *Soc. Sci. Comput. Rev.* , 22 (1), 111-127
- Crawford, E., Kay, J., & McCreath, E. (2002, January 13-16, 2002). *An Intelligent Interface for Sorting Electronic Mail*. Paper presented at the IUI'02 International Conference on Intelligent User Interfaces, San Francisco, California, USA.
- de Vaus, D. A. (2001). *Research Design in Social Research*. London, UK: Sage Publications.
- Deutschens, E., Ruyter, K. d., Wetzels, M., & Oosterveld, P. (2004). Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study. *Marketing Letters*, 15(1), 21-36.
- Dictionary.com Unabridged. (v1.1). Retrieved December 9, 2008, from Dictionary.com website: <http://dictionary.reference.com/>
- Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). New York: Wiley.
- Dillman, D. A., & Bowker, D. K. (2001). The Web Questionnaire Challenge to Survey Methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science*. Lengerich, Germany: Pabst Science.
- Dix, A., Wilkinson, J., & Ramduny, D. (1998, September). *Redefining organisational memory: Artifacts, and the Distribution and Coordination of Work*. Paper presented at the Workshop on Understanding Work and Designing Artefacts, York.
- Dourish, P., Edwards, W. K., LaMarca, A., & Salisbury, M. (1999a). Presto: An Experimental Architecture for Fluid Interactive Document Spaces. *ACM Transactions on Computer-Human Interaction*, 6(2), 133-161.
- Dourish, P., Edwards, W. K., LaMarca, A., & Salisbury, M. (1999b). *Using Properties for Uniform Interaction in the Presto Document System*. Paper presented at the UIST'99 Symposium on User Interface Software and Technology, Asheville, North Carolina, USA.
- Drucker, P. (1959). *Landmarks of tomorrow*. New York: Harper Collins.
- Ducheneaut, N., & Bellotti, V. (2001). E-mail as Habitat: An Exploration of Embedded Personal Information Management. *Interactions*, 8(5), 30-38.
- Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. C. (2003, July 28 - August 1, 2003). *Stuff I've Seen: A System for Personal Information Retrieval and Re-Use*. Paper presented at the SIGIR'03 Conference on Research and Development in Information Retrieval, Toronto, Canada.

- Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: a review of the literature. *International Journal of Information Management*, 20(1), 17-28.
- Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework* (Summary Report for Air Force Office of Scientific Research). Menlo Park, CA: Stanford Research Institute. Document Number)
- Eysenck, M. W., & Keane, M. T. (1990). *Cognitive Psychology: A Student's Handbook*: Lawrence Erlbaum.
- Faichney, J., & Gonzalez, R. (2001). Goldleaf Hierarchical Document Browser. *Australian Computer Science Communications*, 23(5), 13-20.
- Farhoomand, A. F., & Drury, D. H. (2002). Managerial Information Overload. *Communications of the ACM*, 45(10), 127-131.
- Faught, K. S., Whitten, D., & Green Jr., K. W. (2004). Doing Survey Research on the Internet: Yes, Timing Does Matter. *Journal of Computer Information Systems*, 44(3), 26-35.
- Fertig, S., Freeman, E., & Gelernter, D. (1996a). "Finding and Reminding" Reconsidered. *SIGCHI Bulletin*, 28(1), 66-69.
- Fertig, S., Freeman, E., & Gelernter, D. (1996b, April 13-18, 1996). *Lifestreams: An Alternative to the Desktop Metaphor*. Paper presented at the CHI'96 Conference on Human Factors in Computing Systems, Vancouver, Canada.
- Fink, A. (1995a). *How to Ask Survey Questions*. Thousand Oaks: Sage Publications.
- Fink, A. (1995b). *How to Design Surveys*. Thousand Oaks: Sage Publications.
- Freeman, E., & Fertig, S. (1995, November 1995). *Lifestreams: Organizing your electronic life*. Paper presented at the AIII Fall Symposium: AI Applications in Knowledge Navigation and Retrieval, Cambridge, Massachusetts, USA.
- Freeman, E., & Gelernter, D. (1996). Lifestreams: A Storage Model for Personal Data. *SIGMOD Bulletin*, 25(1), 80-86.
- Frohlich, D., & Perry, M. (1994). *The Paperful Office Paradox* (Technical report No. HPL-94-20). Bristol: Hewlett Packard Laboratories. Document Number)
- Gifford, D. K., Jouvelot, P., Sheldon, M. A., & O'Toole, J. W., Jr. (1991, October 1991). *Semantic File Systems*. Paper presented at the Symposium on Operating Systems Principles, Pacific Grove, California, USA.
- Gonçalves, D., & Jorge, J. A. (2004, Jan 13-16, 2004). *Describing Documents: What Can Users Tell Us?* Paper presented at the IUI'04 International Conference on Intelligent User Interfaces, Madeira, Funchal, Portugal.
- Google Inc. (2004). Google Announces Desktop Search. *Press Release*. Retrieved September 9, 2005, from <http://www.google.com/press/pressrel/desktopsearch.html>
- Gwizdka, J. (2000, April 1-6, 2000). *Timely Reminders: A Case Study of Temporal Guidance in PIM and Email Tools Usage*. Paper presented at the CHI'2000 Conference on Human Factors in Computing Systems, The Hague, Netherlands.
- Gwizdka, J. (2004, April 5-8, 2004). *Email Task Management Styles*. Paper presented at the CHI'2004 Conference on Human Factors in Computing Systems Extended Abstracts, Vienna, Austria.
- Hair, J. F., Jr, Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis* (5th ed.). Upper Saddle River, New Jersey, USA: Prentice Hall.
- Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the Flow in Web Site Search. *Communications of the ACM*, 45(9), 42-49.
- Ho, R. (2006). *Handbook of Univariate and Multivariate Data Analysis and Interpretation with SPSS*. Boca Raton: Chapman & Hall/CRC.
- Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7(2), 174-196.

REFERENCES

- Järvelin, K. (2003). Definition of Information Retrieval. In J. Feather & R. P. Sturges (Eds.), *International Encyclopedia of Information and Library Science* (2nd ed., pp. 688). New York: Routledge.
- Jenkins, A. M. (1985). Research Methodologies and MIS research. In E. Mumford, R. Hirschheim, G. Fitzgerald & A. T. Wood-Harper (Eds.), *Research Methods in Information Systems* (pp. 103-117). New York, USA: North Holland.
- Joinson, A. N., & Reips, U.-D. (2005). Personalized salutation, power of sender and response rates to Web-based surveys. *Computers in Human Behavior, In Press, Corrected Proof*.
- Jones, W., & Ross, B. (2006). Human cognition and personal information management. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay & M. T. H. Chi (Eds.), *Handbook of Applied Cognition*. Chichester: John Wiley & Sons.
- Kao, A., Quach, L., Poteet, S., & Woods, S. (2003). *User assisted text classification and knowledge management*. Paper presented at the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA, USA.
- Kaplan, B., & Duchon, D. (1988). Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study. *MIS Quarterly*, 12(4), 570-587.
- Karger, D. R., Bakshi, K., Huynh, D., Quan, D., & Sinha, V. (2003). *Haystack: A Customizable General-Purpose Information Management Tool for End Users of Semistructured Data*. Paper presented at the CIDR'2003 Conference on Innovative Data Systems Research, Asilomar, California, USA.
- Karger, D. R., & Quan, D. (2004, April 5-8, 2004). *Haystack: A User Interface for Creating, Browsing, and Organizing Arbitrary Semistructured Information*. Paper presented at the CHI'2004 Conference on Human Factors in Computing Systems Extended Abstracts, Vienna, Austria.
- Kidd, A. (1994, April 24-28, 1994). *The Marks are on the Knowledge Worker*. Paper presented at the CHI'94 Conference on Human Factors in Computing Systems, Boston, Massachusetts, USA.
- Kitchenham, B., & Pfleeger, S. L. (2002). Principles of Survey Research Part 4: Questionnaire Evaluation. *SIGSOFT Softw. Eng. Notes*, 27(3), 20-23.
- Kwasnik, B. H. (1989). *How a Personal Document's Intended Use or Purpose Affects its Classification in an Office*. Paper presented at the SIGIR'89 Conference on Research and Development in Information Retrieval, Cambridge, Massachusetts, USA.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago, USA: The University of Chicago Press.
- Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1), 55-66.
- Litwin, M. S. (1995). *How to Measure Survey Reliability and Validity*. Thousand Oaks: Sage Publications.
- MacInnis, P. (2003). Desktop democracy. *Computing Canada*, 29(13), 10.
- Mackay, W. E. (1988). *More than just a communication system: diversity in the use of electronic mail*. Paper presented at the CSCW'88 Conference on Computer-Supported Cooperative Work, Portland, Oregon, USA.
- Malone, T. W. (1983). How do people organize their desks? Implications for the design of office information systems. *ACM Transactions on Office Information Systems*, 1(1), 99-112.
- Mander, R., Salomon, G., & Wong, Y. Y. (1992, May 3-7, 1992). *A 'Pile' Metaphor for Supporting Casual Organization of Information*. Paper presented at the CHI'92 Conference on Human Factors in Computing Systems, Monterey, California, USA.
- Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *The Psychological Review*, 63(2), 81-97.
- Mingers, J. (2001). Combining IS Research Methods: Towards a Pluralist Methodology. *Information Systems Research*, 12(3), 240-260.

- Mock, K. (2001, September 9-12, 2001). *An Experimental Framework for Email Categorization and Management*. Paper presented at the SIGIR'01 Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA.
- Nardi, B. A., & Barreau, D. K. (1997). "Finding and Reminding" Revisited: Appropriate Metaphors for File Organization at the Desktop. *SIGCHI Bulletin*, 29(1).
- NetApplications Ltd. (2008). Operating System Market Share. Retrieved December 8, 2008, from <http://marketshare.hitslink.com/operating-system-market-share.aspx?qprid=8>
- Nielsen, J. (2001, August 5, 2001). First Rule of Usability? Don't Listen to Users. *Alertbox*. Retrieved November 11, 2005, from <http://www.useit.com/alertbox/20010805.html>
- Nunamaker, J. F., Chen, M., & Purdin, T. D. M. (1990). Systems Development in Information Systems Research. *Journal of Management Information Systems*, 7(3), 89-106.
- Orlikowski, W. J., & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research*, 2(1), 1-28.
- Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly*, 39(4), 541-574.
- Petter, S. C., & Gallivan, M. J. (2004, January 5-8, 2004). *Toward a Framework for Classifying and Guiding Mixed Method Research in Information Systems*. Paper presented at the HICSS'37 Hawaii International Conference on System Sciences, Hawaii, USA.
- Pinsonneault, A., & Kraemer, K. L. (1993). Survey Research Methodology in Management Information Systems: An Assessment. *Journal of Management Information Systems*, 10(2), 75-105.
- Porter, S. R. (2004). Raising Response Rates: What Works? *New Directions for Institutional Research*, 121, 5-21.
- Price, G. (2004). A New Player in Desktop Search. Retrieved September 15, 2005, from SearchEngineWatch website: <http://searchenginewatch.com/3401711>
- Punch, K. F. (2003). *Survey Research: The Basics*. London, UK: Sage Publications.
- Rao, R., Card, S. K., Johnson, W., Klotz, L., & Trigg, R. H. (1994). *Protofoil: Storing and Finding the Information Worker's paper Documents in an Electronic File Cabinet*. Paper presented at the CHI'94 Conference on Human Factors in Computing Systems, Boston, Massachusetts.
- Rekimoto, J. (1999a). *Time Machine Computing: A time-centric approach for the information environment*. Paper presented at the UIST'99 Symposium on User Interface Software and Technology, Asheville, North Carolina, USA.
- Rekimoto, J. (1999b, May 15-20, 1999). *TimeScape: A time-machine for the desktop environment*. Paper presented at the CHI'99 Conference on Human Factors in Computing Systems Extended Abstracts, Pittsburgh, Pennsylvania, USA.
- Richards, L. (2005). *Handling Qualitative Data: A Practical Guide*. London, UK: Sage Publications.
- Robertson, G. G., Card, S. K., & Mackinlay, J. D. (1993). Information Visualization Using 3D Interactive Animation. *Communications of the ACM*, 36(4), 57-71.
- Robertson, G. G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., & van Dantzich, M. (1998). *Data Mountain: Using Spatial Memory for Document Management*. Paper presented at the UIST'98 Symposium on User Interface Software and Technology, San Francisco, California, USA.
- Rogers, Y. (2004). New Theoretical Approaches for Human-Computer Interaction. *ARIST: Annual Review of Information Science and Technology*, 38(1), 87-143.
- Rosch, E., Mervis, C., Gray, W., Johnson, D., & Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382-439.
- Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B. H., Cai, J., & Liu, X. (2001). *Genre-based navigation on the Web*. Paper presented at the HICSS-34, Maui, Hawaii, USA.

REFERENCES

- Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). *The cost structure of sensemaking*. Paper presented at the Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems. from <http://doi.acm.org/10.1145/169059.169209>
- Schacter, D. L. (1999). The Seven Sins of Memory: Insights From Psychology and Cognitive Neuroscience. *American Psychologist*, 54(3), 182-203.
- Schwarz, N. (1999). Self-Reports: How the Questions Shape the Answers. *American Psychologist*, 54(2), 93-105.
- Schwarz, N., & Oyserman, D. (2001). Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction. *American Journal of Evaluation*, 22(2), 127-161.
- Segal, R. B., & Kephart, J. O. (1999, May 1-5, 1999). *MailCat: An Intelligent Assistant for Organizing E-Mail*. Paper presented at the Agents'99 3rd International Conference on Autonomous Agents, Seattle, Washington, USA.
- Segal, R. B., & Kephart, J. O. (2000, June 29 - July 2, 2000). *Incremental Learning in SwiftFile*. Paper presented at the ICML'2000 17th International Conference on Machine Learning, Stanford, California, USA.
- Sekaran, U. (2000). *Research Methods for Business: A Skill-Building Approach* (3 ed.). New York, USA: John Wiley & Sons.
- Simon, H. A. (1997). The future of information systems. *Annals of Operations Research*, 71(0), 3-14.
- Sutton, M. J. D. (1996). *Document Management for the Enterprise: Principles, Techniques and Applications*. New York: John Wiley & Sons.
- Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, Massachusetts: MIT Press.
- Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004, April 5-8, 2004). *The Perfect Search Engine Is Not Enough: A Study of Orienteering Behaviour in Directed Search*. Paper presented at the CHI'2004 Conference on Human Factors in Computing Systems, Vienna, Austria.
- The American Heritage® Dictionary of the English Language. (2004). Fourth Edition. Retrieved December 9, 2008, from Dictionary.com website: <http://dictionary.reference.com/>
- Treglown, M. (2000, December 8, 2001). *Filing, Piling, Grabbing and Trashing: Applying a Contemporary Theory of Metaphor to User Interface Design*. Paper presented at the PCHCI'01 Panhellenic Conference on Human-Computer Interaction, Patras, Greece.
- Trigg, R. H., Blomberg, J., & Suchman, L. A. (1999, 12-16 September 1999). *Moving document collections online: the evolution of a shared repository*. Paper presented at the European Conference on Computer-Supported Cooperative Work, Copenhagen, Denmark.
- Trochim, W. M. (2002). The Research Methods Knowledge Base. 2nd. Retrieved November 3, 2005, 2005, from <http://trochim.human.cornell.edu/kb/index.htm>
- Trouteaud, A. R. (2004). How you ask counts: a test of internet-related components of response rates to a web-based survey. *Social Science Computer Review*, 22(3), 385-392.
- Whittaker, S., & Hirschberg, J. (2001). The Character, Value, and Management of Personal Paper Archives. *ACM Transactions on Computer-Human Interaction*, 8(2), 150-170.
- Whittaker, S., & Sidner, C. (1996, April 13-18, 1996). *Email Overload: exploring personal information management of email*. Paper presented at the CHI'96 Conference on Human Factors in Computing Systems, Vancouver, Canada.
- Whittaker, S., Terveen, L., & Nardi, B. A. (2000). Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI. *Human Computer Interaction*, 15, 75-106.
- Wilson, T. (2003). Definition of Information Management. In J. Feather & R. P. Sturges (Eds.), *International Encyclopedia of Information and Library Science* (2nd ed., pp. 688). New York: Routledge.

- Xerox. (2006). PARC Innovation Milestones. Retrieved 2006, July 18, from <http://www.parc.com/about/history/>
- Yin, R. K. (1989). *Case Study Research, Design and Methods* (1st ed.): Sage Publications.
- Yin, R. K. (2003). *Case Study Research: Design and Methods* (3rd ed.). Thousand Oaks: Sage Publications.

A. INTERVIEW CONSENT FORM



**THE UNIVERSITY OF AUCKLAND
BUSINESS SCHOOL**
**MANAGEMENT SCIENCE AND
INFORMATION SYSTEMS**

**DEPARTMENT OF MANAGEMENT
SCIENCE & INFORMATION SYSTEMS**

First floor, Old Choral Hall
7 Symonds Street
Auckland, New Zealand
Telephone 64 9 373 7599 ext. 87154
Facsimile 64 9 3737430

The University of Auckland
Private Bag 92019
Auckland, New Zealand

CONSENT FORM FOR PARTICIPANT

THIS CONSENT FORM WILL BE HELD FOR A PERIOD OF SIX YEARS

Title: Personal Information Management: An investigation of how personal information is managed by knowledge workers

Researcher: Ms Sarah Henderson

- I have been given and have understood an explanation of this research project. I have had an opportunity to ask questions and have them answered.
- I understand that I may withdraw myself or any information traceable to me at any time up to one month from the date of the interview without giving a reason.
- I agree to be interviewed for the purposes of this research.
- I agree / do not agree that Information Structure Analyser software may be run on my computer to collect information about folder structures and file types.
- I agree / do not agree that the interview will be audio taped, and understand that, even if I do agree, I may choose to have the recorder turned off at any time.
 - I understand that if I have agreed to be interviewed, I may request to view and amend the transcripts of the interview.
 - I understand that if I have agreed to be interviewed, the tapes will be heard by a transcriptionist. I understand that the transcriptionist will sign a confidentiality agreement ensuring the confidentiality of my information.

Signed:

Name:
(please print clearly)

Date:

**APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN SUBJECTS
ETHICS COMMITTEE**

on for a period of years, from/...../.....
Reference/.....

B. INTERVIEW RECORD SHEET

Personal Digital Document Management			
<h3 style="margin: 0;">DATA SUMMARY</h3> <div style="display: flex; justify-content: space-between; align-items: flex-start; margin-top: 10px;"> <div style="width: 60%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Participant Code:</div> </div> <div style="width: 35%; text-align: right;"> DATE : <div style="display: flex; justify-content: space-around; width: 100px;"> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> </div> </div> </div>			
<h3 style="margin: 0;">DEMOGRAPHICS:</h3> <div style="margin-top: 10px;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Job Title:</div> <div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 15%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Age:</div> </div> <div style="width: 15%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Gender: M F</div> </div> <div style="width: 20%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Years in Current Position:</div> </div> <div style="width: 20%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Years in Current Field:</div> </div> </div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> How well organised overall: 1 2 3 4 5 1=Disorganised 5=Very organised </div> <div style="border: 1px solid black; height: 60px; margin-top: 10px;">Main tasks & activities:</div> </div>			
<h3 style="margin: 0;">COMPUTER EXPERIENCE:</h3> <div style="display: flex; justify-content: space-between; align-items: flex-start; margin-top: 10px;"> <div style="width: 20%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Years of Exp:</div> </div> <div style="width: 20%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Number of Computers:</div> </div> <div style="width: 60%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Proficiency: Beginner Intermediate Advanced Expert</div> </div> </div> <div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 30%;"> <p>Storage Media Used</p> <div style="margin-top: 5px;"> <input type="checkbox"/> Local Hard Drive <input type="checkbox"/> Network Drive <input type="checkbox"/> Floppy Drive <input type="checkbox"/> Optical Drives <input type="checkbox"/> USB Drives <input type="checkbox"/> Cecil </div> </div> <div style="width: 30%;"> <p>Computer Types Used</p> <div style="margin-top: 5px;"> <input type="checkbox"/> Desktop PC <input type="checkbox"/> Laptop PC <input type="checkbox"/> PDA <input type="checkbox"/> TabletPC </div> <p style="font-size: small; margin-top: 5px;">Indicate primary computer</p> </div> <div style="width: 40%;"> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;">Years with this computer:</div> <div style="border: 1px solid black; height: 25px; margin-bottom: 10px;">Last reimage:</div> </div> </div> <div style="margin-top: 10px;"> <p>Main work area: <input type="checkbox"/> My Documents <input type="checkbox"/> Desktop <input type="checkbox"/> Both <input type="checkbox"/> Other</p> <p>File Explorer: <input type="checkbox"/> Windows Explorer <input type="checkbox"/> My Computer <input type="checkbox"/> Both <input type="checkbox"/> Other</p> <p>Primary Navigation: <input type="checkbox"/> Folder windows <input type="checkbox"/> Tree view</p> <p style="margin-left: 20px;"> <input type="checkbox"/> Keyboard <input type="checkbox"/> All <input type="checkbox"/> Other </p> </div>			

C. INTERVIEW PROTOCOL

Personal Digital Document Management

INTERVIEW GUIDE

1. Demographics & Computer Use

Use Data Summary Sheet

2. Overall

Thinking about your file system and your documents on your primary computer now, how **well organised** would you say it is?

Prompt (if disorganised): Why do you say that?

Prompt (if organised): What advantages do you think you get from having it organised?

Prompt: does it work well **for you**?

Prompt: is the **time** worth it?

Prompt: is it **important**?

Now, imagine **Microsoft** are **redesigning** the way Windows allows you to manage your documents and folders and the Desktop.

What is the **best** thing about it, which you would like to **keep**?

What is the **worst** thing about it, which you could like to **remove**?

What is the **most** useful new thing they could **add**?

3. Desktop & Docs

What do you keep on your **desktop**?

Prompt: permanent organisation?

Prompt: who created these documents?

Prompt: spatial?

Prompt: longevity.

Prompt: edit or reference

Prompt: destination

Prompt: why

Prompt: shortcuts

Where do you keep the rest of your documents?

What sort of things do you create folders for?

Prompt: reorganise?

Prompt: number of items as a trigger?

4. Creating Documents

How do you create new documents? (ask for demo if possible)

Prompt: when do you name them?

Prompt: naming schemes

Prompt: rename other people's docs?

Prompt: include names of folders?

5. Locating

How would you open a document you used **within 2 days**?

Prompt: success/failure rate?

How would you open a document you used **within 1 year**?

Prompt: success/failure rate?

Prompt: **total failure**? Why?

Do you use the **Search** facility?

Prompt: for your files or others?

Prompt: what search criteria?

How do you **view** your document folders?

Prompt: tree?

Prompt: new window?

Prompt: back and forward?

Prompt: up one level?

Prompt: view?

Prompt: sorting?

6. Versions, Delete & Backup

Do you keep separate files for different **versions**?

Prompt: how do you distinguish?

Prompt: lose track?

Do you accidentally have **multiple copies**?

Prompt: why?

Prompt: a problem?

Do you **delete** documents?

Prompt: why for a single file?

Prompt: recycle bin?

Prompt: disk space?

What happens to **old documents**?

Prompt: backup procedures

7. Working Practices

Multiple documents open for editing?

Save email attachments?

Save files on web pages?

Prompt: special location?

How have your management practices **changed through time**?

D. INTERVIEW TRANSCRIPT EXTRACT

Could you show me where you keep your documents?

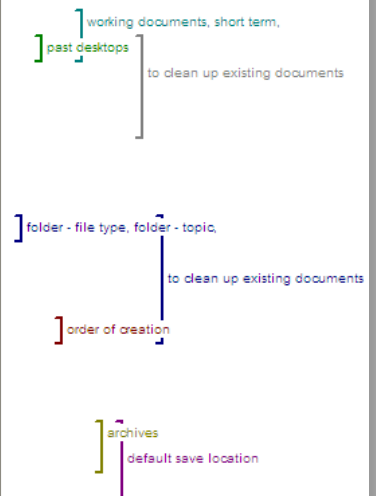
Well, pretty much anything that I'm using right now or have been using recently are the files on the Desktop. Periodically these files get dumped into folders on the Desktop. Periodically those folders get dumped into, I guess other folders either in My Documents or at times onto the C: drive. Like I'll create a folder which will have stuff on the first of the first two thousand and something [01012004]. Which means I'll pretty much start again with a clean slate.

So when you take all the stuff on your desktop and put it into folders, how do you name the folders?

Some of the folders are set by category, either by topic or file type. So I've got like, some of the folders are set up for "Word Docs", "Excel Docs", "Access Docs". But generally if that happens then, I mean like, all the Access stuff is related to the same topic anyway. Other stuff is stuck into folders which are topic related, like Lectures, ok, so they'll consist of PowerPoints and Word docs and maybe Visio files or ERDs. [Icons are ordered by creation date]

And when you move those into My Documents, do you keep the same kind of structure in there?

That's sort of hard to say. The stuff in My Documents is generally just files, again, which were used at one point. Some of the folders in here such as, like the general My Documents, sometimes I just save there by default.



E. CONTACT NOTE

Contact Note of Interview with Participant F

Major concepts

Repeating group of subfolders. Uses the same subfolder structure for each course. Has even developed tools to automatically create the structure for him.

Synchronisation a major problem. Uses a utility to keep files synchronised between a desktop, a laptop and two network folders.

Uses shortcuts to folders to avoid duplication.

Naming schemes duplicate folder information to allow for document to be seen in other contexts.

Renames file downloaded because they have different significance.

Creates parallel structures to maintain correspondence between emails and docs, and between install files and installed applications.

Questions Raised

Are many people this obsessive about keeping to a rigid structure?

He is happy with the tools provided because there is the ability to write his own scripts using VBA, VBS and batch files to augment the system to work as he desires it. Are many people able to do this?

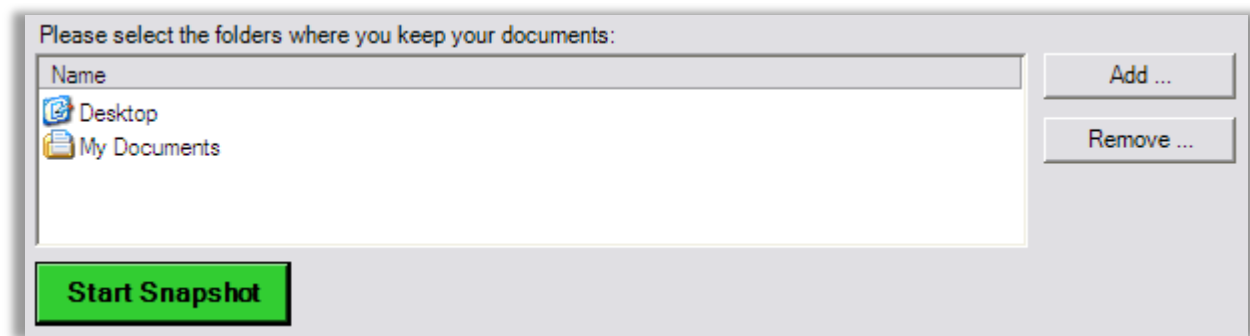
Requirements

A document should keep its context information.

But, document context is personal. Perhaps need a document to have multiple contexts. An author context and an owner context?

F. FILE SYSTEM SNAPSHOT TOOL

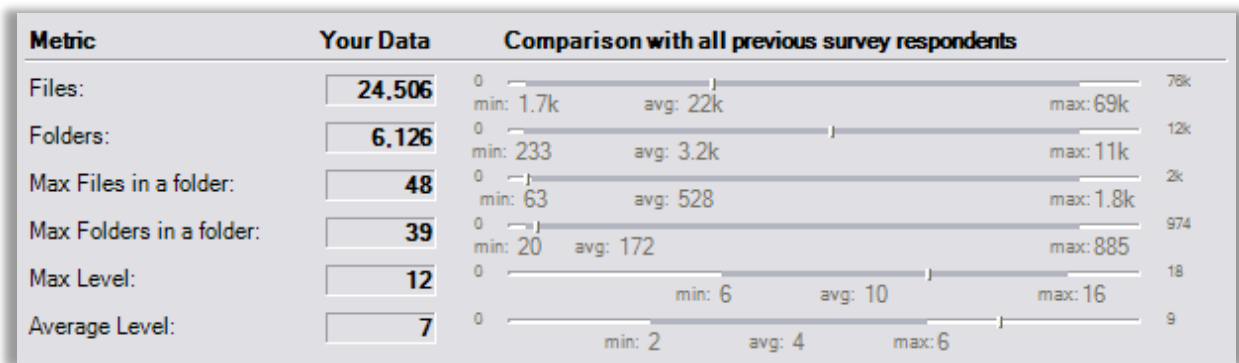
The file system snapshot tool was written in VB.NET using the Microsoft .NET Framework. It made use of the built-in IO library to access and traverses the file system to record the structure of the files and folders it contains. The system was designed to be extremely easy to use and to encourage the user to complete the snapshot. The system includes by default the two most common document storage locations (Desktop and My Documents folder) and makes it very easy for the user to add and remove folders from the snapshot. The large green ‘Start Snapshot’ button makes it clear to users how to proceed.



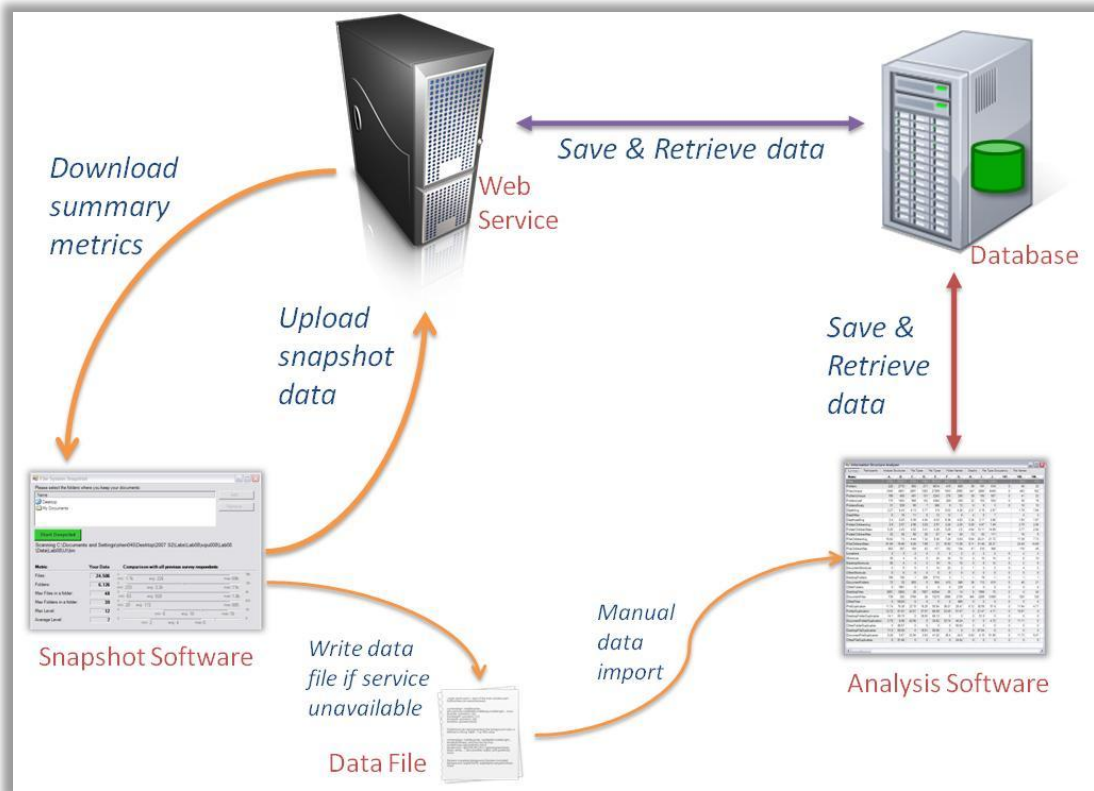
When the application first loads, it connects to a central web service that stores summary information about previous snapshots and downloads the latest summary data. It displays this summary to the user in a panel and indicates that if they perform the snapshot, they'll see how they compare to others. This is designed to encourage them to complete the snapshot, as people are curious about how their file system compares to that of other people.

Metric	Your Data	Comparison with all previous survey respondents			
Files:	0	0			76k
		min: 1.7k	avg: 22k	max: 69k	
Folders:	0	0			12k
		min: 233	avg: 3.2k	max: 11k	
Max Files in a folder:	0	0			2k
		min: 63	avg: 528	max: 1.8k	
Max Folders in a folder:	0	0			974
		min: 20	avg: 172	max: 885	
Max Level:	0	0			18
			min: 6	avg: 10	max: 16
Average Level:	0	0			7
			min: 2	avg: 4	max: 6

As the snapshot proceeds, this data is updated live and the user can see how their filesystem compares. This is designed both to provide feedback during the execution of the snapshot and to prevent the user from cancelling the snapshot. Depending on the size of the file system, the snapshot can take up to 15 minutes to complete. The summary statistics are updated as each file and folder is encountered in the snapshot.



The information extracted about each folder and file was the name, path, and creation time, access time and modification time. Relationships between folders and files were recorded by assigning every single directory and document encountered a unique identifier, and using this to track parent and child relationships. The information was buffered and uploaded to the web service in batches of 100 records. Each execution of the snapshot generated a unique identifier so that all the records from the same snapshot were related to each other. The web services recorded the data from all the snapshots in a database. If the web service was unavailable, the information was instead written to a text file to be uploaded to the database at a later point. The application notifies the user if this is the case and asks the user to email the file to the researcher. The diagram below shows the overall architecture of this solution.



G. SURVEY QUESTIONS

Attitudes

For all the questions in this survey, please think about the documents and folders stored on the hard drive of your primary work computer.

1 How organised would you say your documents and folders usually are?

- ☐ Very organised (1)
- ☐ Somewhat organised (2)
- ☐ Not very organised (3)
- ☐ Not at all organised (4)

2 How much do you agree or disagree with each of the following statements?

	Strongly disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly agree (5)
I feel my documents are well organised to suit my working habits	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I sometimes get annoyed at the time taken to locate my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that someone else would be easily able to find things in my system	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I sometimes wish there was a better way to organise my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am quite happy with the way I manage my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would rather spend less time organising my documents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The time I spend organising my documents is worth it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would be embarrassed to show someone how my documents are organised	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think it is important to have my documents well organised	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If someone showed me a better way to organise my documents, I would probably change the way I do it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

New System Features

Imagine Microsoft are changing the way Windows allows you to manage your documents and folders (including both the file system and the Desktop), and they have asked you for your opinion.

3 What would you tell them was the **best** thing about the current system, the one you would definitely like to **keep**?

4 What would you tell them was the **worst** thing about the current system, the one you would definitely like to **remove**?

- 5** What would you tell them was the most useful **new** thing they could add to the system?

Desktop

- 6** What is displayed on your Desktop?

- ☐ The default background (1)
- ☐ A single colour (2)
- ☐ A picture (3)
- ☐ A pattern (4)
- ☐ Rotating wallpaper (5)
- ☐ "Active Desktop" content (6)

- 7** Do you usually keep documents on your Desktop? *Answer no if the only things you usually store on the Desktop are shortcuts. Answer yes if you usually store documents in folders on your Desktop*

- ☐ Yes (1)
- ☐ No (2)
- ☐ No, because I didn't know I could (3)

- 8** Are the documents on your Desktop organised into folders?

- ☐ Yes (1)
- ☐ No (2)

- 9** Who do you think created most of the documents on your Desktop?

- ☐ Me (1)
- ☐ Other people (2)

- 10** Do you sometimes move documents to specific areas on your Desktop (e.g., top left)?

- ☐ Yes (1)
- ☐ No (2)
- ☐ I didn't know I could (3)

- 11** How long do documents tend to stay on your Desktop?

- ☐ Days (1)
- ☐ Weeks (2)
- ☐ Months (3)
- ☐ Years (4)

- 12** Are the documents on your Desktop mainly for reference, or mainly for editing?

- ☐ Mainly for reference (1)
- ☐ Mainly for editing (2)
- ☐ Other: (3)

13 When a document leaves your Desktop, where is it most likely to go?

- ☐ To a permanent location in a folder (1)
- ☐ I send it to someone then delete it (2)
- ☐ To an archive location (3)
- ☐ Deleted (4)
- ☐ Other: (5)

14 What is the main reason for putting documents on your Desktop?

- ☐ It is a convenient place to save things (1)
- ☐ It is easy to access frequently used documents (2)
- ☐ It reminds me of things I need to do (3)
- ☐ Other: (4)

My Documents

15 Do you use the system created My Documents folder to store your documents?

- ☐ Yes (1)
- ☐ No (2)

16 (If No) Why not?

- ☐ Must be stored elsewhere to be backed up (1)
- ☐ Must be stored elsewhere so they will not get backed up (2)
- ☐ Old habit of storing elsewhere (3)
- ☐ I don't know (4)
- ☐ Other: (5)

Creating and Naming Documents

17 How do you usually create new documents?

- ☐ I open an existing document and use Save As (1)
- ☐ I copy and rename an existing document in the file system (2)
- ☐ I right-click inside a folder and choose New (3)
- ☐ I use the Quick Start or Office Toolbar (4)
- ☐ I use the New Office Document from the Start Menu (5)
- ☐ I open the application (6)

18 When you usually give your documents a name?

- ☐ Before I create any content (1)
- ☐ After I have created some content (2)

19 Do you sometimes rename other people's documents in your file system?

- ☐ I usually rename them (1)
- ☐ I'll rename them only if the name is not meaningful to me (2)
- ☐ I usually just save them with the name the author gave them (3)
- ☐ I don't have other people's documents in my file system (4)

20 How often do you use naming schemes or patterns in your document names?

- ☐ Always (1)
☐ Often (2)
☐ Seldom (3)
☐ Never (4)

21 Do you sometimes include the name of one of the parent folders in the name of the document?

- ☐ Yes (1)
☐ No (2)

22 (If Yes), what is the main reason you do this?

23 What is the minimum information you usually need to know about one of your own documents in order to accurately guess its contents?

- ☐ File name only (1)
☐ File name and file type (2)
☐ File name and its location in my folders (3)
☐ File name and file type and its location in my folders (4)
☐ Something else: (5)

Creating and Naming Folders

24 When do you usually create folders?

- ☐ I create folders before I have files to put in them (1)
☐ I create folders as I need to put documents into them (2)
☐ I create folders after I have a number of files that need to be organised (3)

25 How important or unimportant is it to you to be able to organise your documents by each of these dimensions?

	Very unimportant (1)	Unimportant (2)	Neutral (3)	Important (4)	Very important (5)
Time Periods (e.g. year or semester)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Projects or Courses	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
File Types (e.g. Excel, Zipped)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Subject or Topic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Purpose or Use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Locating Documents

26 How would you open a document you last opened within the past 2 days? If you use multiple methods, choose the one you would usually use first.

- ☐ I know exactly which folder it is in and browse the folder then open the document directly (1)
- ☐ I know approximately which folder it is in and browse the likely folders until I see it (2)
- ☐ I open the application and use the Recent menu (3)
- ☐ I use the Start > Recent documents (4)
- ☐ I use the Find Files function (5)
- ☐ I use some other type of search (6)
- ☐ Other: (7)

27 How often are you able to locate the document this way on the first try?

- ☐ Always (1)
- ☐ Often (2)
- ☐ Sometimes (3)
- ☐ Seldom (4)
- ☐ Never (5)

28 How would you open a document you last opened over 1 year ago? If you use multiple methods, choose the one you would usually use first

- ☐ I know exactly which folder it is in and open it directly (1)
- ☐ I know approximately which folder it is in and browse the likely folders until I see it (2)
- ☐ I use the Windows Search for Files and Folders function (or similar) (3)
- ☐ I use some other type of search. (4)

29 How often are you able to locate the document this way on the first try?

- ☐ Always (1)
- ☐ Often (2)
- ☐ Sometimes (3)
- ☐ Seldom (4)
- ☐ Never (5)

Searching

30 How do you usually use a search facility to locate documents on your computer? A search facility could be the Windows Search for Files and Folders function or another search tool such as Copernic or Google Desktop

- ☐ It is the first method I will use (1)
- ☐ I will try other ways first, but if I don't find it quickly, I will use a search facility (2)
- ☐ I will try other ways first, and will only use a search facility as a last resort (3)
- ☐ I will try other ways, and will never use a search facility (4)

31 When you use a search facility, what are you normally looking for?

- ☐ My own documents (1)
- ☐ System files (2)
- ☐ Files from elsewhere (3)

32 How frequently or infrequently do you use each of these options when using a search facility?

	Never (1)	Seldom (2)	Sometimes (3)	Often (4)	Always (5)
All or part of the file name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Keywords or phrases from the file contents	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
File type or extension	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
When the file was created or modified	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Size of the file	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

33 Have you ever been completely unable to find a file (by any means)?

- ☐ Yes (1)
☐ No (2)

34 (If Yes) What do you think is the most usual reason for this?

- ☐ The file is probably on my computer but I can't remember enough about it to locate it (1)
☐ The file was never saved on this computer but probably somewhere else (2)
☐ The file was once on this computer, but is now probably somewhere else (3)
☐ The file was deleted (4)
☐ The file never existed at all (5)
☐ Other: (6)

Viewing Documents**35 Answer each of the following questions about the way you usually view your documents inside your folders**

	Yes (1)	No (2)
Do you usually have the tree visible?	<input type="radio"/>	<input type="radio"/>
Do you usually use the tree to navigate?	<input type="radio"/>	<input type="radio"/>
Do you usually have the address bar visible?	<input type="radio"/>	<input type="radio"/>
Do you usually use the address bar to navigate?	<input type="radio"/>	<input type="radio"/>
Do you usually use the back and forward button on the toolbar to navigate?	<input type="radio"/>	<input type="radio"/>
Do you usually use the 'Up one level' button on the toolbar to navigate?	<input type="radio"/>	<input type="radio"/>
Do you usually use the keyboard to navigate?	<input type="radio"/>	<input type="radio"/>
Do you usually open each folder in a new window?	<input type="radio"/>	<input type="radio"/>

36 Which view do you usually use to display your files?

- ☐ Large Icons (1)
☐ Small Icons (2)
☐ List (3)
☐ Details (4)
☐ I didn't know there were options (5)

- 37 (If Details)** How frequently or infrequently do you sort your documents by each of these dimensions?

	Never (1)	Seldom (2)	Sometimes (3)	Often (4)	Always (5)
Name	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Size	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
File Type	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Date Modified	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Creating and Naming Documents

- 38** Do you sometimes have separate files for different versions of a document?

- ☐ Yes (1)
☐ No (2)

- 39** How do you usually distinguish between the different versions?

- ☐ I put version numbers in the file names (1)
☐ I put dates in the file names (2)
☐ I put descriptions in the file names (3)
☐ I put the files into another folder (4)
☐ Other: (5)

- 40** Do you sometimes lose track of which file is the most recent version of a document?

- ☐ Yes (1)
☐ No (2)

- 41 (If Yes)** Why do think this is?

Copies

- 42** Do you sometimes accidentally have more than one copy of the same document on your computer (the exact same document, not just a different version)?

- ☐ Yes (1)
☐ No (2)

- 43 (If Yes)** What is the usual reason why this happens?

- ☐ I forgot I had the document and created it again (1)
☐ I couldn't find the old document and created it again (2)
☐ I saved it from email more than once (3)
☐ I downloaded it from the internet more than once (4)
☐ Other: (5)

Deleting and Backup

44 What is the most common reason why you would delete a file?

- ☐ I have another copy somewhere else (1)
- ☐ I have a more recent version (2)
- ☐ I could easily get this document again (3)
- ☐ The document was temporary, with no lasting value (4)
- ☐ To save disk space (5)
- ☐ I would never delete a file (6)
- ☐ Other: (7)

45 Do you usually use the Recycle Bin or permanently delete?

- ☐ I delete into the Recycle Bin (1)
- ☐ I delete into the Recycle Bin, but I empty the Recycle Bin frequently (2)
- ☐ I permanently delete (3)

46 (If you use the Recycle Bin) How often do you usually retrieve documents from the Recycle Bin?

- ☐ Daily (1)
- ☐ Weekly (2)
- ☐ Monthly (3)
- ☐ Yearly (4)
- ☐ Never (5)

47 When a folder and its contents are no longer actively used, what is most likely to happen to it?

- ☐ I will delete it (1)
- ☐ I will make sure there is a backup somewhere else (e.g., on CD) then delete it (2)
- ☐ I will leave it where it is (3)
- ☐ I will move it to a different location but keep it on disk (4)
- ☐ I will add it to a zip file (or similar) and delete it (5)
- ☐ Other: (6)

48 Do you usually take hard drive space into account when deciding whether to delete or keep documents?

- ☐ Yes (1)
- ☐ No (2)

Demographics

49 How experienced would you consider yourself at using Windows?

- ☐ Beginner (1)
- ☐ Intermediate (2)
- ☐ Advanced (3)
- ☐ Expert (4)

50 How many computers (including PDAs) do you store personal documents on?

- ☐ One (1)
- ☐ Two (2)
- ☐ Three or four (3)
- ☐ Five to seven (4)
- ☐ More than seven (5)

51 How many years have you been working with computers as part of your work?

years

52 How many years have you been doing the same type of work you currently do?

years

53 What is your age?

- ☐ < 20 (1)
- ☐ 20-29 (2)
- ☐ 30-39 (3)
- ☐ 40-49 (4)
- ☐ 50-59 (5)
- ☐ 60-69 (6)
- ☐ > 70 (7)

54 What is your gender?

- ☐ Male (1)
- ☐ Female (2)

55 Are you general staff or academic staff?

- ☐ General (1)
- ☐ Academic (2)

56 What department or unit are you primarily associated with?

- ☐ Accounting & Finance (1)
- ☐ Commercial Law (2)
- ☐ Computer Services Unit (3)
- ☐ Economics (4)
- ☐ Faculty Office (5)
- ☐ Graduate School of Business (6)
- ☐ Information Systems & Operations Management (7)
- ☐ International Business (8)
- ☐ Management & Employment Relations (9)
- ☐ Marketing (10)
- ☐ Property (11)
- ☐ Tamaki Division (12)
- ☐ TechSite Services (13)
- ☐ Other (14)

- 57** **What is your job title?** *Leave this empty if you feel that answering would compromise your anonymity.*

Comments

- 58** **Is there anything else about your personal document management that you think I should know?**

Continue > > >